A User Study on Public Health Events Detected within the Medical Ecosystem

Avaré Stewart
L3S Research Center
Hanover, Germany 30167
Email:stewart@L3S.de

Eelco Herder
L3S Research Center
Hanover, Germany 30167
Email: herder@L3S.de

Matthew Smith
Leibniz University Hannover
Hanover, Germany 30159
Email: smith@rvs.uni-hannover.de

Wolfgang Nejdl
L3S Research Center
Hanover, Germany 30167
Email: nedjl@L3S.de

***Abstract*—The great influx of Medical-Web data makes the task of computer-assisted gathering and interpretation of Social Media-based Epidemic Intelligence (SM-EI) a very challenging one. State-of-the-art approaches usually use supervised machine learning algorithms to gather data from a variety of sources in this medical ecosystem, mining this data for specific event patterns and information discovery. Supervised approaches not only limit the type of detectable events, but also requires learning examples be given to the machine learning algorithm in advance. On the other hand, the more generic and flexible unsupervised machine learning methods currently produce such complex results, that the domain experts are not capable of assessing the results in a natural and efficient manner.**

In this paper, we present a novel framework with which SM-EI field practitioners can interact with medical ecosystem data, and assess the results of such complex unsupervised SM-EI algorithms. The assessment framework and the unsupervised epidemic event detection algorithm have been fully implemented and a quantitative study is presented to show the validity of the new approach to SM-EI.

I. Introduction

The multi-disciplinary area of Social Media-Based Epidemic Intelligence (SM-EI) has emerged as a type of computer-assisted intelligence gathering that is devoted to harnessing information in the Medical Ecosystem for the purpose of detecting public health events, from unstructured text. The information in the digital ecosystem is exploited by epidemic investigators when tackling the dual task of: i) monitoring the prevalence of known public health events; and ii) detecting emerging ones, as quickly as possible.

The Medical Ecosystem constitutes a social media data space, in which the interests of the participants are specifically devoted to medical and health issues. The Medical Ecosystem ranges from physicians and nurses, patients and relatives to health officials and average users. By incorporating social media from this medical ecosystem, as a new source of information, SM-EI systems augment the traditional public health event monitoring tasks, in such as way that the new information will clarify and not cloud, the situational awareness of an epidemic investigator.

The first step in realizing SM-EI, is the detection of public health events. This is typically accomplished by relying upon two different types of statistical detection methods: supervised [9], [6], [8] and unsupervised detection [7], [2], [1]. Both approaches have their strengths (in terms of the type of surveillance required) and trade offs (in terms of the complexity of its inputs and outputs).

The more common, supervised approach, tunes a learner to detect specific types of information, and generally solves a binary classification problem to determine the relevance of sentences or documents, for public health events. Although this allows known types of public health events to be detected with a good level of confidence, it relies upon human effort in creating a large amount of learning examples and feature engineering to train the supervised algorithm.

Given the the dual-detection task in SM-EI, and the limitations of a supervised approach, the less well studied, but more generic and flexible, unsupervised approach (e.g., those based on clustering) is considered. Since unsupervised approaches detect events based on inherent patterns that exist in a data collection, they are capable of detecting new events. Unsupervised approaches also have the advantage that the domain expert need not specify the learning examples in advance.

A. Problem Statement

Detecting public health events in an unsupervised manner leads to very complex results, which pose a significant challenge for an epidemic investigator, given the number of potential clusters. Additionally, since the pattern is not labeled apriori, the significance and meaning of the pattern must be interpreted. In order to ensure that the unsupervised methods produce results that are valuable for the human users, it is crucial that SM-EI systems not only detect public health events, but also consider a user-centric approach which emphasizes both: an assessment of the public health event quality, and representation.

The problems to be addressed are: 1) detecting patterns meaningful as epidemic events within the Medical Ecosystem, in an unsupervised manner; 2) presenting the detected patterns to a human effectively; thereby allowing SM-EI domain experts to easily interpret any epidemic patterns that are mined from the ecosystem; and 3) enabling SM-EI domain experts to analyze the epidemic intelligence data.

To address these problems, we present a novel framework with which SM-EI field practitioners can interact with the social media data and assess the results of complex unsupervised SM-EI algorithms. The **user-centric** pattern assessment framework constitutes: *pattern mining*, *pattern pruning* and a

ISBN: 978-1-4577-0872-5 (c) 2011 IEEE

user-centric *pattern evaluation* involving field practitioners. A key aspect of the framework is a feedback interaction loop which involves; tuning a system to better help these users in their epidemic investigation activities.

B. Contribution and Significance of the Work

This paper makes several contributions in the context of information discovery within a Medical Ecosystem for epidemic intelligence gathering and analysis. The significance of these contributions is in addressing the challenges within the field of Social Media-Based Epidemic Intelligence. Since the amount of social media data is quite high and its content, noisy, poor decisions taken using social media may lead to serious consequences or wasted investigative effort. Social Media-Based Epidemic Intelligence systems are of little benefit if practitioners can not trust the public health events they produce. Our user-centric assessment framework, the natural language epidemic intelligence mining algorithm and the unsupervised epidemic event detection algorithm have been fully implemented and a quantitative study is presented to show the validity of the new approach to SM-EI.

The paper is structured as follows. In Section II, the user-centric framework and its evaluation method for unsupervised pattern assessment is introduced. Then, in the Section III, the experimental setting is described, and the results of the experiments are presented. In Section IV, the related work is presented. The paper ends with conclusions in Section V and remarks on future work.

II. Field Practitioner-Assisted Assessment

In this section, we present our framework for the Practitioner-Assisted assessment of unsupervised public health events. An overview of the framework is depicted in Fig. 1 and each stage is presented in the discussion that follows.

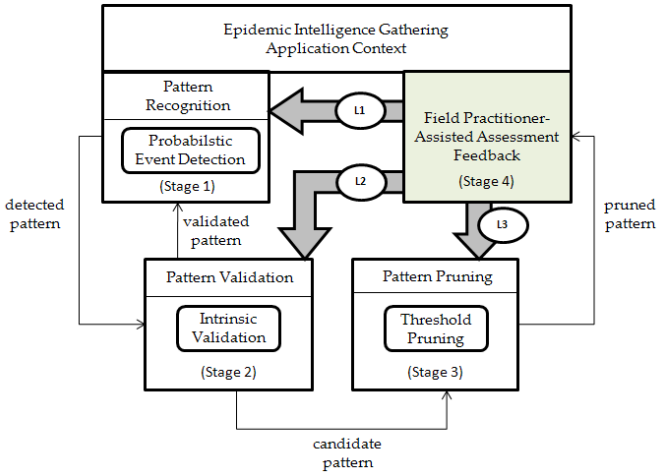


Fig. 1: Assessment Framework Overview

A. Pattern Recognition

We consider an unsupervised public health event, to be a type of pattern that is recognizable by a *Pattern Recognition*, and a selected event detection algorithm. The type of pattern we consider is a clustering of documents. We define an unsupervised event, more formally:

Definition 1: Unsupervised Event: An unsupervised event, $\mathcal{I}_j$, is considered an unlabeled class, c_j, that represents a clustering of documents, which has been detected by an unsupervised pattern recognition engine, Φ_K. The engine is a function that produces a weight matrix, $w_{ij} \in W_d$, such that each document d_i, where i = $\{1,...N\}$, is assigned to every unlabeled class, c_j, where j = $\{1,..,K\}$, with some weight $w_{ij} \geq 0$; where N is the total number of documents, and K is the total number of patterns.

Probabilistic Event Detection, in Fig. 1, represents different probabilistic algorithms that can be used for public health event detection. If the terms are also weighted then the event detection produces a weight matrix, $w_{jk} \in W_t$, such that each term, t_k, is assigned to every cluster, c_j, with some weight $w_{jk} \geq 0$; where where k = $\{1,...M\}$ and M is the total number of terms. The patterns produced by the *Pattern Recognition* are then used as input for the *Pattern Validation* stage.

B. Pattern Validation

Pattern Validation allows the different parameters of a pattern recognition algorithm to be iteratively tuned. For example, in unsupervised event detection, the number of classes, K, is required to be given as input. A value for K can determined using one of several intrinsic cluster validation metrics [3]. Different types of validations, other than the choice of K, may be required. The type of validation is determined based on the choice of the event detection algorithm and task.

C. Pattern Pruning

As noted in Definition 1, many patterns may be produced, since the pattern recognition engine associates each document with every class, with some weight. In practice, associating every documents to a class will not provide meaningful results for everyday tasks, so some of the candidate patterns produced by the *Pattern Validation* stage, are eliminated before being presented to the user. Candidate patterns are pruned at three levels of granularity: term-level and document-level (intra-pattern); and cluster level (intra-pattern). The definitions for inter- and intra- pattern pruning are given in Definitions 2 and 3, respectively.

Definition 2: Inter-Pattern Pruning: A pattern, I_j, is pruned if $Quality(\mathcal{I}_j) \leq \alpha$, for some given quality measure α.

Definition 3: Intra-Pattern Pruning: Given a weight matrix, W_d, and a cluster, $\mathcal{I}_j$; a document d_i , is pruned if $w_{ij} \leq \beta$. Likewise, a term is pruned from the cluster if $w_{jk} \leq \gamma$, where $w_{jk} \in W_t$.

In Definitions 2 and 3, the $Quality(\mathcal{I}_j)$, and values for α, β and γ are chosen according to task and algorithm-specific criteria.

D. Practitioner-Assisted Feedback

The *Field Practitioner-Assisted Assessment* stage takes pruned patterns as input for the user to assess their quality.

ISBN: 978-1-4577-0872-5 (c) 2011 IEEE

The feedback interaction loops (denoted by $L1$, $L2$, $L3$ in Fig. 1) signify that: the features of the algorithm in the *Pattern Recognition* stage ($L1$), the validation technique in the *Pattern Validation* stage ($L2$), or the pruning criteria in the *Pattern Pruning* stage ($L3$) are all subject to adaptation, based on feedback from the user assessment. This is intended to improve the quality of public health event, as well as the documents and words that make up a cluster. In Section III, we describe the experimental results when applying this framework using a selected pattern recognition algorithm for the task of SM-EI.

III. Experiments

The objectives of the experiments is threefold. First, we are interested in a field practitioner assessment on the quality of public health events that have been detected in an unsupervised manner, given our choice for an intrinsic validation metric, and threshold values of α, β and γ for the pruning criteria. We examine if different combination of weights properly group documents into logical clusters, and whether these clusters are meaningful to the users. Second, we are interested in knowing whether the representation of the patterns as word clouds are useful to the users. Third, we seek to learn how the results of the user assessment would serve as a feedback interaction loop to influence the pattern mining, intrinsic cluster validation, and pattern pruning processes.

A. Data Set

As a dataset, medical blogs from MedWorm, a moderated blog medical blog aggregation, were collected via RSS during an eight month period from May 2009 through January 2010. This period is known to coincide with the 2009 Swine Flu pandemic. We used this set, so that users could have some familiarity with the events detected. In total, 30,822 blogs (186,230 sentences) were selected by retrieving documents from the subcategories under the heading of *infectious disease*. Since the RSS text contained only summaries, the urls from the RSS were used to crawl the website. The raw html was processed by stripping all boilerplate and markup code using the method introduced by Kohlschütter et.al. [4]. The final data set then contained only unstructured text.

B. Experimental Setting

Pattern Recognition Algorithm: For the unsupervised event detection, we adapted a Retrospective Event Detection (RED) [5] algorithm. This model for event detection, is extended to include medical conditions, which is important for SM-EI, since it allows a public health event to be defined in terms of the entity types: *medical condition* and *location*. We extracted the entities using OpenCalais[1]. Of the 30,822 documents, 2,532 documents (27,900 sentences) contained medical conditions or locations.

The patterns produced by this algorithm consist of a set of probabilities, which we use as weights; these weights consist of: 1) a document weight matrix (W_d); 2) a term weight matrix (W_t); and 3) a set of weights for each detected cluster.

[1]http://www.opencalais.com

Pattern Validation: In order to determine the number of clusters to use as input for the algorithm, we run the pattern detection algorithm for values a $K = \{10...100\}$ and selected the K for which the cluster cohesion was the highest, this corresponded to a value of K = 93. Since the detection algorithm uses a random initialization, the results vary for each trial. We selected the best trial by running the algorithm for 100 trials and selected the trial having the highest log-likelihood.

Pattern Pruning: The 93 candidate patterns, were pruned by using intra-patten and inter-patten pruning. The threshold values for pruning the clusters (α), documents (β), terms (γ) were selected based on a quartile distribution of their probabilities.

In order to determine the impact of using quartiles as a pruning criteria, we selected two different quartile ranges for the documents and clusters. We used the notation L and H to correspond to the range of probabilities in the first quartile (low probability values), and the fourth quartile (high probability values), respectively. Using a combinations of these quartiles, three pruning criteria: HH, HL, and LH, were used to prune the clusters and documents. The ordering: HH, HL and LH, reflects the increasing noise, with respect to the probability values. All term probabilities were taken from the H quartile range and in total, 9 of the 93 clusters (3 clusters for each pruning criteria), were presented to the users for evaluation.

C. Case Study Methodology

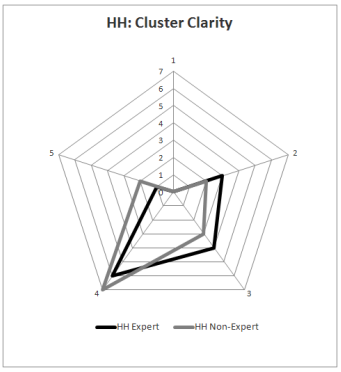
User Groups: Two user groups were asked to assess the quality of the 9 clusters, and their representation as word clouds. The first group consisted of five practitioners in the field of epidemiology. We label this group as the "Expert" group. The second group, as a basis for comparison, consisted of non-practitioners, five individuals with backgrounds in the area of user centered design; we label this group as the "Non-Expert".

D. Clustering Clarity

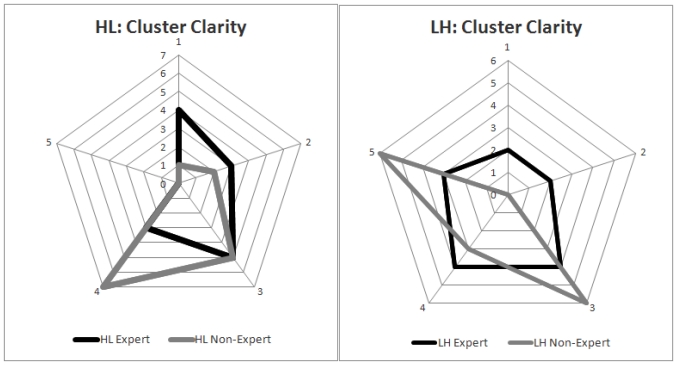
We were interested in knowing whether the high probability clusters correspond to clusters that actually make sense to the users. In order to do this, each cluster was evaluated by the user for its overall clarity. The users were asked the extent to which the set of documents in the group made sense to them, using a scale of: 1=confusing, 5=clear. Fig. 2, shows the frequency of each rating, for each pruning criteria among each user group. Quite noticeable for the HH-pruning criteria, is that the shapes of the two graphs for the experts and non-experts are quite similar. Both user groups found the clusters within the HH-pruning criteria to be rather clear (see Fig. 2a).

In contrast, for the LH-pruning criteria in Fig. 2c, the experts seem to be quite neutral and do not rate strongly for, or against the clarity of the clusters. This suggests that in the presence of lower probability clusters, in the LH category, the experts are more unclear about what constitutes a pattern, unlike the non-expert users.

ISBN: 978-1-4577-0872-5 (c) 2011 IEEE



(a) HH



(b) HL (c) LH

Fig. 2: Overall Clarity for Pruning Criteria HH (2a); HL (2b); and LH (2c) based on the extent to which the set of documents for the group makes sense to the user; using the scale: 1=confusing,5=clear.

For the HL-pruning criteria (Fig. 2b) one can notice that there is a clear overlapping (that splits around the most neutral rating of 3). This suggests that given the mixture of low probability documents with high probability clusters, the users are completely mixed, about the clarity of the clusters.

Finally, we also notice in the Figs. 2b and 2c, that in the presence of increasing noise, the expert users are more conservative about their ratings, unlike the non-experts, who still rate the clarity of the cluster comparatively higher (in the range of $3 \cdots 5$), than the experts. The Expert group, assesses the quality of the clusters differently than the Non-Expert group in the presence of more noise (i.e., lower probability).

E. Document Fit within a Cluster

In Section III-D, the clusters were evaluated by the users for its overall clarity. In this experiment, we were interested in assessing the quality of clusters, with respect to the documents contained therein. The users were asked to rate the extent to which the five documents in each of the three pruning criteria fit the cluster. Fig. 3, 4 and 5 represent the results of the percent for trial 1 using the HH, HL and LH pruning criteria, respectively.

In Fig. 3, we notice that the percent agreement is is stratified across the values of (1.0, 0.6 and 0.4) for both user groups.

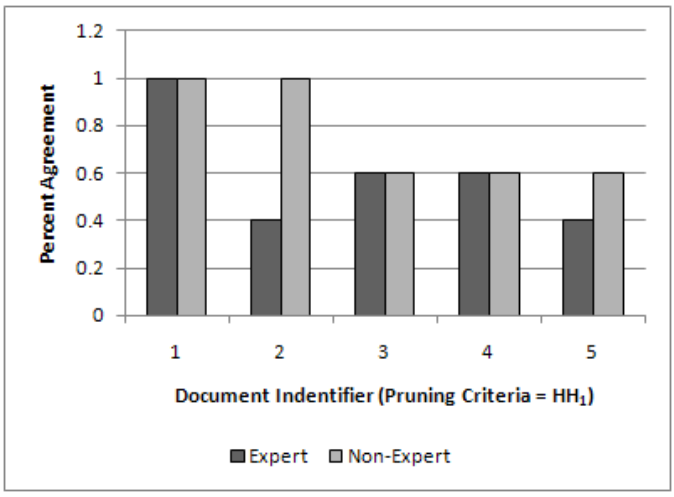


Fig. 3: Percent agreement for the extent to which the documents of the HH pruning criterial fit the cluster.

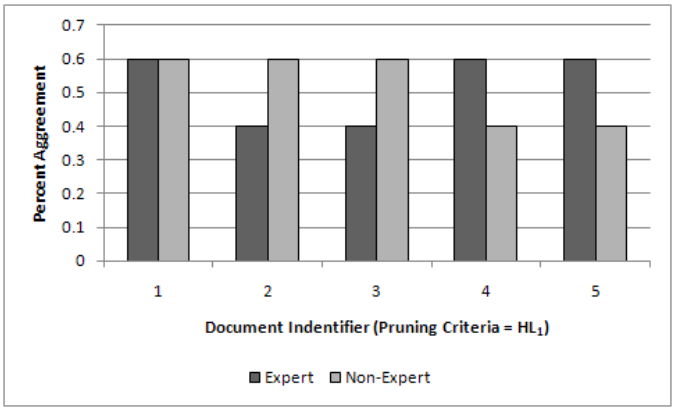


Fig. 4: Percent agreement for the extent to which the documents of the HL pruning criterial fit the cluster.

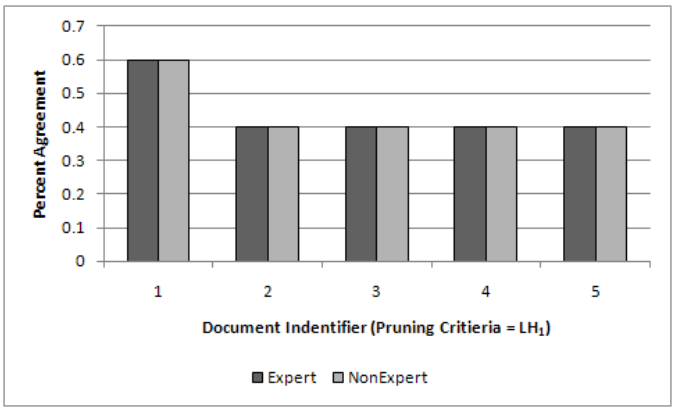


Fig. 5: Percent agreement for the extent to which the documents of the LH pruning criterial fit the cluster.

This stratification suggests that users saw distinct sub-clusters of documents as belonging to the same equivalence class. In Fig. 4, we see a similar stratification. In contrast, however to the HH criteria, here we see that the peak stratification is

significantly lower: at a value of 0.6 compared with 1.0 in Fig. 3. This suggests that, again, users perceive sub-clusters, yet, they were less confident about the meaningfulness of the document with respect to the cluster. This can be explained by the fact that the documents that were included, were taken from the low probability range of the quartile. This means that low probability documents do, in fact, lead to less meaningful and less clear clusters from the users perspective.

Quite remarkably, in Fig. 5, we notice that for nearly four out of the five documents, both user groups, show the least agreement about the fit of the document within the cluster. This lead us to believe that when using high probability documents with a low probability cluster, the users do not perceive that the documents fit the clusters well.

In our instantiation of the framework, we relied up the cohesion of the documents with the cluster to validate the cluster. We believe that alternative metrics should be considered. The silhouette metric, for example, takes into account, not only the cohesion of the document within the cluster, but also its separation, with respect to the other clusters. Validating the clusters according to other criteria, would bring further insights into the task of assigning documents to a cluster.

Practitioner Assisted Feedback Loop (L2): *Based on the stratification in the percent agreement scores we propose that another cohesion metric for intrinsic validation in Pattern Validation stage, be considered.*

We conclude that the HH Pruning Criteria is meaningful for the users, whereas the HL and LH clusters are less clear and meaningful.

F. Cluster Representation

In order to ensure that the unsupervised methods produce results that are valuable for the human users, it is crucial that SM-EI systems also provide a means to represent the clusters so users can interpret them with ease. We consider a cluster representation based on two different types of word clouds and seek to know which on the users find useful, if any. Fig. 6 shows the results when users, were asked to choose which type of word cloud they thought best represents the cluster.

The results overwhelmingly show that the users preferred the term frequency word cloud over the named entity word cloud. The term frequency clouds was constructed using the frequency for all terms that were present in the documents for, the given cluster. In contrast, the named entity word clouds were constructed by using the named entities with a threshold probability as determined by the fourth quartile of the term probability distribution.

Practitioner Assisted Feedback Loop (L2): *Since the term frequency word cloud was preferred to the named entity word cloud, we propose that the choice of features for the event detection algorithm should not only contain entity types such as medical condition and location, but also include non-entity terms.*

G. User Remark and Feedback

In this section, qualitative feedback based on the users remarks are presented. All users found the word clouds

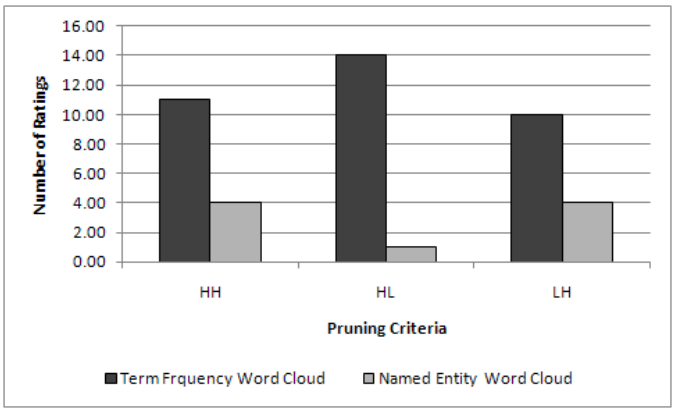


Fig. 6: Number of Ratings indicating the word cloud that users thought best describes the set of documents for the group. The choice of words cloud representations where: Term Frequency and Named.

effective. However, when asked what other representations would meet their needs, they thought it useful to have more robust and interactive word clouds. Some user's comments for more interactivity, include: a) the use of color coding to identify different entity types in the cloud; b) interactive clouds to remove words and redraw word clouds; c) the use of an ontology to collapse semantically related terms that are in the cloud; and d) the elimination of very small cloud terms.

Practitioner Assisted Feedback Loop (L3): *Since the users wanted to eliminate very small terms and collapse semantically related terms: we propose that: 1) more pruning should be done for the term pruning phase, and 2) terms that are semantically related, should be aggregated before they are used as input into the pattern recognition algorithm.*

Although the experts in the study are not trained in the use of Web 2.0 tools, they were quite clear about the ways in which the system can help them to better manage and manipulate the complexity of the outputs for unsupervised public health events, in a Web 2.0 style.

Practitioner Assisted Feedback Loop (L1): *Based on the desire for the users to have more control over the terms that appear in the word cloud, more interactivity should be produced to help users manipulate and digest the content.*

Finally, users thought other types of representations could be considered. These representations dealt mostly with alternative ways to see the same underlying data. Users even suggested having a toggle button, so that they could decide which alternate representation, they could see, for a given situation.

In summary our results show that:

1) Field practitioners can identify clear cluster that have been produced by an unsupervised event detection algorithm,
2) patterns with a high cluster probability and high document probability are better suited for field practitioners, in digesting and interpreting the meaning of the pattern,
3) the use of term frequency word clouds can help field practitioners to distinguish patterns with respect to their

ISBN: 978-1-4577-0872-5 (c) 2011 IEEE

quality.

IV. Related Work

Our work differs significantly from the existing work done in detecting public health event, in generatl and unsupervised public health events , specifically, since other systems, do not include domain experts to assess the quality of the resulting clusters. Furthermore, our work is based on an user-centric approach to SM-EI. More specifically, our view of assessment emphasizes a domain expert's perspective - that is orthogonal to the notion of relevance, precision or recall as presented in other SM-EI assessment systems [8], [7], [2]. Particularly for the problem domain of Epidemic Intelligence, field practitioners indeed provide the best insight on the quality of a detected pattern.

To this end, we combine an intrinsic cluster validation, with a user assessment, whereby the results of the user evaluation serve as a feedback loop which, in turn, influence the pattern mining, intrinsic cluster validation, and pattern pruning processes. The novelty of our work is in understanding how field practitioners value and interpret the results produced by a probabilistic event detection algorithm for their work in intelligence gathering.

V. Conclusions and Future Work

In this paper, we presented a novel framework with which SM-EI field practitioners can assess the results harnessed from the Medical Ecosystem for intelligence gathering to detect public health events from unstructured text. interact with public health indicators data and asses the results of complex unsupervised SM-EI algorithms. We presented formalizations for characterizing public health events and how this is embedded in the user-centric SM-EI framework. We presented a study including over 30,000 blog entries and numerous SM-EI domain experts to validate the both the components and the underlying unsupervised SM-EI event detection algorithm and the feedback interaction loop between the two.

We have shown that 1) field practitioners are able to find clear clusters that have been produced by an unsupervised event detection algorithm, 2) patterns with a high cluster probability and high document probability are better suited for field practitioners, in digesting and interpreting the meaning of the pattern and 3) the use of term frequency word clouds can help field practitioners to distinguish patterns with respect to their quality.

The impact of such work in practice and research is two-fold. First, is an improved understanding of the types of visualization and representations that are useful to domain experts in the areas of epidemiology. Second, is bridging the gap between system mining and filtering, and the domain experts in the field, who must rely upon a summarized interpretation and an elucidation of facts for Social Media-Based Epidemic Intelligence systems.

Future work will include incorporating the information from the practitioner assisted feedback loops. We also plan as extension of the social media sources to include, more noisy data, in a streaming setting, to further stress test the ability of our framework to represent this complex data in a human understandable way. Finally an interesting area for future research is to include official sources of information from the Medical Ecosystem as a baseline such as the WHO or ProMed-Mail database statistics, to enable an automated correlation between human-centric SM-EI data and other Medical Ecosystem data.

VI. Acknowledgments

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

References

[1] C. D. Corley, D. J. Cook, A. R. Mikler, and K. P. Singh. Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health*, 7(2):596–615, 2010.

[2] M. Fisichella, A. Stewart, K. Denecke, and W. Nejdl. Unsupervised public health event detection for epidemic intelligence. In *CIKM 2010: 19th ACM Conference on Information and Knowledge Management*, New York, NY, USA, 2010. ACM.

[3] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 7(2/3):107–145, 2010.

[4] C. Kohlschüter, P. Fankhauser, and W. Nejdl. Boilerplate detection using shallow text features. In B. D. Davison, T. Suel, N. Craswell, and B. Liu, editors, *WSDM*, pages 441–450. ACM, 2010.

[5] Z. Li, B. Wang, M. Li, and W.-Y. Ma. A probabilistic model for retrospective news event detection. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 106–113, New York, NY, USA, 2005. ACM.

[6] M. Naughton, N. Stokes, and J. Carthy. Sentence-level event classification in unstructured texts. *Information Retrieval*, 13(2):132–156, April 2010.

[7] A. Stewart, M. Fisichella, K. Denecke, and W. Nejdl. Detecting public health indicators from the web for epidemic intelligence. In *eHealth 2010*, 2010.

[8] P. von Etter, S. Huttunen, A. Vihavainen, M. Vuorinen, and R. Yangarber. Assessment of utility in web mining for the domain of public health. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 29–37, Los Angeles, California, USA, June 2010. Association for Computational Linguistics.

[9] Y. Zhang. *Automatic Extraction of Outbreak Information from News*. PhD thesis, University of Illinois at Chicago, 2008.