[Demo] Integration of text- and web-mining results in EpidVis

Samiha Fadloun[1,2], Arnaud Sallaberry[1,3], Alizé Mercier[4,5], Elena Arsevska[6], Pascal Poncelet[1,2], Mathieu Roche[4,7]

[1] LIRMM, CNRS, Univ. Montpellier, France
[2] Univ. Montpellier, France
[3] Univ. Paul-Valéry Montpellier 3, France
[4] Cirad, Montpellier, France
[5] ASTRE, Univ. Montpellier, Cirad, Inra, Montpellier, France
[6] Institute of Global Health and Epidemiology, United Kingdom
[7] TETIS, Univ. Montpellier, APT, Cirad, CNRS, Irstea, Montpellier, France

Abstract. The new and emerging infectious diseases are an incising threat to countries due to globalisation, movement of passengers and international trade. In order to discover articles of potential importance to infectious disease emergence it is important to mine the Web with an accurate vocabulary. In this paper, we present a new methodology that combines text-mining results and visualisation approach in order to discover associations between hosts and symptoms related to emerging infectious disease outbreaks.

Keywords: visualization, text-mining, query, epidemiology

1 Introduction

Online detection and monitoring of animal disease outbreaks is crucial for disease surveillance and early warning systems. Epidemiologists regularly query web-pages using various formulations to obtain up-to-date information on disease outbreaks, but this task may take several days before finding the sought-after information. Visualization could nevertheless facilitate their web searches as compared to time-consuming traditional searches. This paper presents EpidVis, a new visual web query tool designed for epidemiologists. It consists of several views that help build and launch queries, and visualize the results. Moreover, it supports external information integration to help epidemiologists enrich their knowledge and adapt their queries. This paper focuses on the integration of text- and web-mining results in EpidVis tool.

Most of the current visualization tools in epidemiology domain focus on web results in general [1,2]. They do not deal with storing knowledge and adapting queries. They mainly show web results related to a given disease outbreak [3,4,5]. They use results from different sources (RSS feeds, reports, alerts, etc.) showing them as plots on a map to represent spatial information, or as statistical diagrams to represent aggregated information [6].

To summarize, most of the previous visualization techniques focus on showing the query results. In animal epidemiology, we can distinguish three main categories of keywords (*diseases*, *hosts*, and *symptoms*), and identify relationships. The visualization can help the epidemiologists to express these categories and relationships, a crucial step that is not incorporated in the previous approaches. This visual expression can help the user to build and launch queries.

Section 2 summarizes the EpidVis tool. The suggestion view that integrates text- and web-mining results is described in Section 3. Finally, Section 4 concludes and describes future work.

2 EpidVis tool

Epidemiologists regularly search on the web resources in order to get new insights about disease outbreaks. They use their own knowledge as keywords and manually write and launch the queries on search engines like Google, Bing, Yahoo, etc. They often copy/paste keywords stored in text file. They use three main sets of keywords (categories): *diseases*, *hosts*, and *symptoms*. For example, one can launch the query: *'influenza chicken lethality'*. In this case, *'influenza'* is a *disease*, *'chicken'* is a *host*, and *'lethality'* is a *symptom*. In this context, this is crucial to take into account strong or weak relationships between the keywords of different categories. Epidemiologists also use external knowledge to build queries, e.g. knowledge transmitted from their colleagues, or data extracted with text-mining or statistical approaches [7].

3 The Suggestion View

This section describes how to enrich the keywords by using external knowledge. This external knowledge is provided as a specific file containing keywords and relationships that exist between them. For instance, it contains relationships between *hosts* and *symptoms* associated with a specific *disease*. Our goal is to provide a new view taking into account this external knowledge (i.e. text- and web-mining results) to suggest links between keywords to experts.

In order to measure the relevance of association between *hosts* and *symptoms*, statistical measures have been used like Dice measure. In our context, this measure is defined as the number of times where *hosts* and *symptoms* appear in the same context (in a corpus or in the Web) over the sum of the total number of times that each one appears in the corpus (i.e. text-mining approach) or in the Web (i.e. web-mining approach) for each disease. This approach has been adapted with other statistical measures (i.e. Mutual Information and Cubic Mutual Information). This contribution is detailed in [7]. The values of association measures represent the input of suggestion view of EpidVis.

The suggestion window of EpidVis contains two main views: the suggested keyword view (Figure 1.a) which shows the data coming from the external file,

and the current keyword view (Figure 1.b) which shows the data already available in the keyword manager. Both of the views are based on a circle splitted into arcs representing the keywords. In order to make the circles homogeneous, we plot the union of the keywords of the keyword manager and the suggested keywords. Underlined keywords represent words from the external file in Figure 1.a, and words from the keyword manager in the Figure 1.b. The selected keyword is represented at the top of the circles (blue arc "bluetongue" in the example). The other keywords are visualized around it. Each arc has a name, and a color according to its category. We put a color degradation to make a difference between keywords in the same category. Relations between keywords are represented by curves between the corresponding arcs. Each arc is splitted into sub arcs and the width of a sub arc represents the weight of the links starting from it. Each link has an inverse color interpolation related to categories of related arcs. We hide links between the selected keyword and the other keywords, in order to avoid cluttering: we show them when hovering the mouse on the arc, or selecting them. A slider (Figure 1.c) is provided to filter suggested keywords and relationships by weight of relationships. Text information is also available as shown in Figure 1.d. It notifies to the epidemiologist the actions performed as described in the next section. Each type of action is encoded with specific colors: red color for selected triplets to be added to the keyword manager, purple color for the already added triplets, and black for the triplets to be deleted. Note that our system holds several interactive features to explore the suggested data, and to add some of these suggested data to the keywords manager dataset.

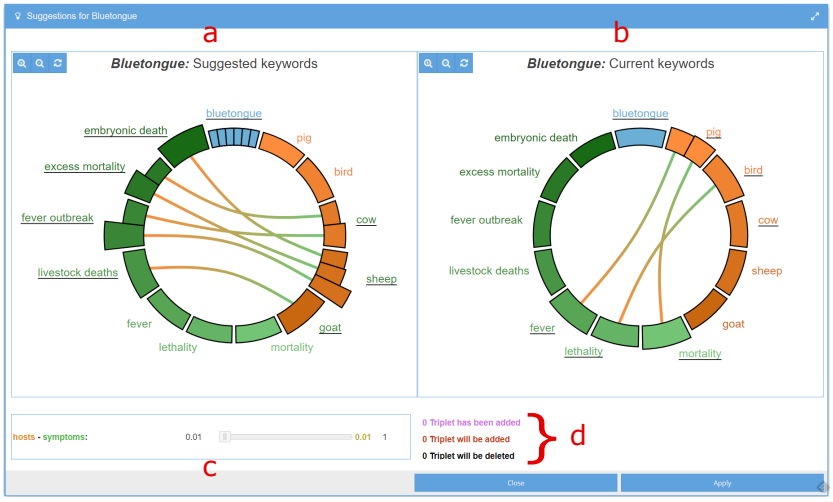

Fig. 1. The suggestion view. **(a)** Relationships and keywords suggested an external file. **(b)** Relationships and keywords already available in the keyword manager. **(c)** Slider to filter relationships and keywords according to their weight, **(d)** Result of different actions.

The usefulness and usability of the different views of EpidVis have been evaluated. A survey was presented to a group of 12 participants containing experts and non experts in epidemiology. It can be noticed that the participants have highly appreciated the suggestion evaluation visualizations, i.e. 8.4/10 for usefulness criterion, and 8.1/10 for usability criterion.

4 Conclusion and Future Work

In this paper, we present EpidVis, a new visual web querying tool for animal disease surveillance. This tool helps epidemiologists to build queries, launch them on the web, and visualize the results. Also, external knowledge can be integrated using the suggestion view that takes into account text- and web-mining results. For future work, we plan to extend our tool, in order to cover other application domains.

Acknowledgments: This work was supported by the Ministry of Higher Education and Scientific Research of Algeria and the SONGES project (FEDER and Occitanie). We thank Renaud Lancelot (ASTRE, Cirad) and Sarah Valentin (ASTRE & TETIS, Cirad) for their expertise in epidemiological surveillance.

References

1. Dörk, M., Carpendale, S., Collins, C., Williamson, C.: Visgets: Coordinated visualizations for web-based information exploration and discovery. IEEE Transactions on Visualization and Computer Graphics **14**(6) (2008) 1205–1212
2. Peltonen, J., Belorustceva, K., Ruotsalo, T.: Topic-relevance map: Visualization for improving search result comprehension. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces, Association for Computing Machinery (2017) 611–622
3. Van den Broeck, W., Gioannini, C., Gonçalves, B., Quaggiotto, M., Colizza, V., Vespignani, A.: The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. BioMed Central infectious diseases **11**(1) (2011) 37
4. Collier, N., Doan, S., Kawazoe, A., Goodwin, R.M., Conway, M., Tateno, Y., Ngo, Q.H., Dien, D., Kawtrakul, A., Takeuchi, K., et al.: Biocaster: detecting public health rumors with a web-based text mining system. Bioinformatics **24**(24) (2008) 2940–2941
5. Freifeld, C.C., Mandl, K.D., Reis, B.Y., Brownstein, J.S.: Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports. Journal of the American Medical Informatics Association **15**(2) (2008) 150–157
6. Neher, R.A., Bedford, T.: nextflu: real-time tracking of seasonal influenza virus evolution in humans. Bioinformatics **31**(21) (2015) 3546–3548
7. Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., Dufour, B.: Identification of associations between clinical signs and hosts to monitor the web for detection of animal disease outbreaks. International Journal of Agricultural and Environmental Information Systems **7**(3) (2016) 1–20