Data and Text Mining in Analysis of the Emergence of H5N1 Virus Strain

Peter Frometa, Systems Engineer, SPSS, Inc.

OBJECTIVE

This paper describes a series of data mining techniques used to gather and analyze and disseminate large amounts of data from numerous sources in English as well as Chinese. The objective of the analysis is to attempt to identify locations where the data may indicate a current or future outbreak of the A-H5N1 strain of the flu virus.

BACKGROUND

The highly pathogenic A-H5N1strain of the flu virus was first described as affecting humans in 1997 in Hong Kong. Until that time, this strain was thought to only infect birds.

Currently, humans can only be affected by the disease through close contact with live infected birds. According to the World Health Organization, as of June 2006, there were 228 cases of the H5N1 strain in humans, leading to 130 deaths[1].

This analysis focuses on a federal health agency in the United States that is concerned with tracking the spread of the virus across various countries. This example, specifically studies the emergence of the virus throughout provinces in China.

Traditionally, China has been slow to admit to the spread of infectious diseases. It is not uncommon for China to report cases of infectious diseases to the World Health Organizations weeks or months after an outbreak.

METHODS

We begin the analysis process by collecting available data provided or made available by the Chinese Health Ministry. This data may contain confirmed and suspected cases of avian flu in humans, as well as cases of various flu-like symptoms reported by local hospitals and regional health care organizations throughout the country. In addition, we collected current data from a wide variety of Chinese web logs, forums and local and regional news sources.

All unstructured (free form text) data collected from various internet sources is automatically translated from simplified Chinese into English. The translated Chinese documents are then made available to a categorization engine. This process automates the tedious process of organizing documents into logical categories. In this example, our health agency is actively involved in tracking numerous infectious diseases and outbreaks throughout the world. One of the many categories created through the taxonomy manager is for the A-H5N1 strain of the bird flu.

Documents likely to be related to this virus strain are identified and made available to a linguistics based text extraction engine. Linguistics-based text mining is based on the field of study known as *natural language processing* (NLP), also known as *computational linguistics*.

Concepts determined to be linked to potential cases of avian flu are automatically normalized, linked to their source documents, and made available for data manipulation and analysis.

The extracted concepts are merged with available databases and logs of hospital symptoms and confirmed or suspected cases of avian flu. The data is then passed to an association algorithm. Association rules associate a particular conclusion with a set of conditions. In this case, the conclusion or consequent of interest is a specific province or city within China. The set of conditions that we provide as possible links to the consequent consist of all available data including, category descriptors from the categorization engine, extracted concepts from web logs, forums and news feeds, as well as the available structured database tables from hospitals and the Ministry of Health.

RESULTS

A sample of the resulting association rulesets can be interpreted as follows:
Web log chatter that included the concept "flu", category descriptors of "poultry products" and "poultry industry", reported hospital cases of a history of fever, and reported respiratory symptoms were all associated with the province of Taiyuan.

The above antecedents occurred in 2.21% of the data set. Over 79% of the above antecedents led to an association with the province of Taiyuan.

CONCLUSIONS

From the above results an investigator may conclude that the reported hospital symptoms in Taiyuan are likely to be related to the H5N1 virus even if the presence of the virus has yet to be confirmed. The use of data mining and text mining algorithms may provide this agency actionable information weeks or months before confirmation of an outbreak in a specific location by government authorities.

REFERENCES

1 http://en.wikipedia.org/wiki/Global_spread_of_H5N1