PROCEEDINGS OF SCIENCE

Propagation Routes Analysis of HPAI Outbreaks using Sequential Pattern Mining

Zhenshun Xu[1], **Heesu Choi, Jonguk Lee**
Korea University Sejong Campus , 2511 Sejong-ro, Sejong city, 30019, Korea
E-mail: jssoon77@korea.ac.kr;chlgmltn420@korea.ac.kr; eastwest9@korea.ac.kr

Daihee Park[2 3]
Korea University Sejong Campus , 2511 Sejong-ro, Sejong city, 30019, Korea
E-mail: dhpark@korea.ac.kr

Yongwha Chung[3]
Korea University Sejong Campus , 2511 Sejong-ro, Sejong city, 30019, Korea
E-mail: ychungy@korea.ac.kr

Highly pathogenic avian influenza (HPAI) virus can spread rapidly, resulting in high mortality and severe economic damage. To minimize the damage incurred from such diseases, it is necessary to develop technology for analysing livestock disease and predicting livestock disease propagation. In this study, we propose a novel big data analytics model using extensive volumes of livestock disease occurrence data accumulated over an extended period. In particular, we describe a sample process based on a specific scenario that elicits information that generates sequential dissemination routes of HPAI outbreaks by applying sequential pattern mining in this paper.

ISCC 2015
18-19, December, 2015
Guangzhou, China

[1]Speaker
[2]Correspongding Author
[3]This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1D1A3A01018731) and BK (Brain Korea) 21 Plus Program.

PoS(ISCC2015)022

1.Introduction

Since the first detection of the HPAI virus at the end of 2003, five outbreaks of HPAI have been reported in South Korea [1]. Highly infectious livestock diseases have increased the threat to human life, causing considerable economic damage and environmental problems. To minimize the damage incurred from such diseases, it is necessary to develop technology for analyzing livestock disease occurrence data. To date, an extensive variety of studies has reported on HPAI outbreaks. Seo et al. used computational fluid dynamics to estimate the dispersion of the virus attached to aerosols produced by livestock[2] . Lee et al. constructed a direct HPAI spread network based on the relationships between farms using poultry-related business data and an indirect HPAI spread network using the aerial spread from each farm during the HPAI outbreak in 2008[3]. Tuncer and Le studied the effects of air travel on the spread of avian influenza from Asian and Australian cities to the United States. Real air travel data was used to model the disease spread by individuals who were susceptible to or were infected with pandemic avian influenza[4].

Unlike the current research perspectives, in this study, we propose an efficient big data analytics solution for the analysis of HPAI in South Korea using accumulated long-term livestock disease occurrence data provided by the Ministry of Agriculture in South Korea [5]. The proposed analysis model provides comprehensive and detailed analysis results utilizing simple on line analytic processing (OLAP) operations and powerful data-mining techniques on a multidimensional data cube. In particular, we focus on describing a sample process based on a specific scenario that elicits information that generates sequential dissemination routes of HPAI outbreaks by applying sequential pattern mining in this paper. We test the feasibility and applicability of the proposed HPAI analysis model by implementing a HPAI analysis system and applying it to the analysis of HPAI outbreaks in South Korea.

2.Materials and Methods

2.1 HPAI Outbreak Analysis System

In this section, we introduce an HPAI analysis model that can analyze extensive long-term HPAI data using OLAP operations and powerful data-mining techniques on a multidimensional data cube. Figure 1 illustrates the overall concept of the proposed analysis system.

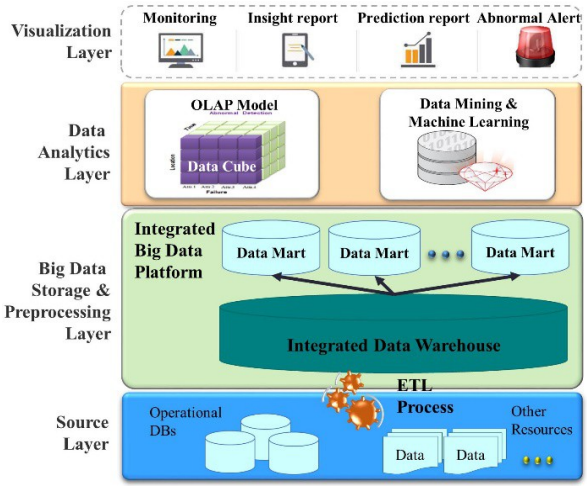

Figure 1: HPAI Outbreak Analysis System Architecture

As can be seen in this figure, the system consists of four layers from data resources to the representation of analysis results: source data repositories, data warehouse and pre-processing, data analytics with OLAP and data mining techniques, and visualization. The various forms of data repositories in the first layer are delivered to the next layer of storage. To ensure that only data in the standard format is stored in the data warehouse, a process called extract, transform, and load (ETL) is performed to preprocess the data collected from the different data sources. In the third layer, we perform multidimensional analysis using OLAP and data-mining techniques. The results of this layer provide valuable information to strongly suggest an improved disinfection policy and strategy for preventing and controlling epidemic diseases in the last visualization layer's interfaces.

2.2 Data Sources

In our experiments, we used the livestock disease occurrence data provided by the Ministry of Agriculture in South Korea over 13 years from 2003 to 2015 (until May 2015) [5]. The row data contains serveral attributes such as diagnosis time, farm address, farm's name, legal disease grade, livestock disease's name, and variety of livestock. Since the first detection of HPAI virus at the end of 2003, five HPAI occurrences have been reported in South Korea: 2003/2004, 2006/2007, 2008, 2010/2011, and 2014/2015. Table 1 presents a summary of HPAI occurrence status across the country including outbreak year, numbers of outbreak, first outbreak site, and total outbreak sites. Figure 2 illustrates the HPAI outbreaks in South Korea over 13 years, from 2003 to 2015 (until May 2015), graphically.

Year (starting season)	Numbers of outbreak	First outbreak site	Total outbreak sites
2003/2004 (winter)	19	Chungbuk.Eumseong	10
2006/2007 (winter)	7	Jeonbuk.Iksan	5
2008 (spring)	33	Jeonbuk.Gimje	19
2010/2011 (winter)	53	Jeonbuk.Iksan	25
2014/2015 (winter)	361	Jeonbuk.Gochang	5

Table 1: Summary of HPAI Outbreak Status

2.3 Sequential Pattern Mining

In this section, we briefly describe a data-mining techniques to retrieve useful information for a comprehensive and in-depth understanding of HPAI diseases: sequential pattern mining. Sequential pattern mining is the discovery of frequently occurring ordered events or subsequences as patterns. Given a set of sequences, where each sequence consists of a list of events (or elements) and each event consists of a set of items, and a user-specified minimum support threshold, sequential pattern mining identifies all the frequent subsequences; that is, the subsequences whose occurrence frequency in the set of sequences is no less than a minimum support threshold [6]. In this paper we used 'arulesSequence' packages in R with SPADE algorithm [7-8]. The data set for sequence mining consists of a collection of input sequence. Each input sequence in the data set has a unique identifier called TID which it represents the year of occurrence, and each event in a given input sequence also has a unique identifier called EID which it represents the occurrence time in same TID, and each event contains sequence items which it represents the locations of occurrence, in the paper. Table 2 presents an example data set for sequential pattern mining including TID, EID, and Items.

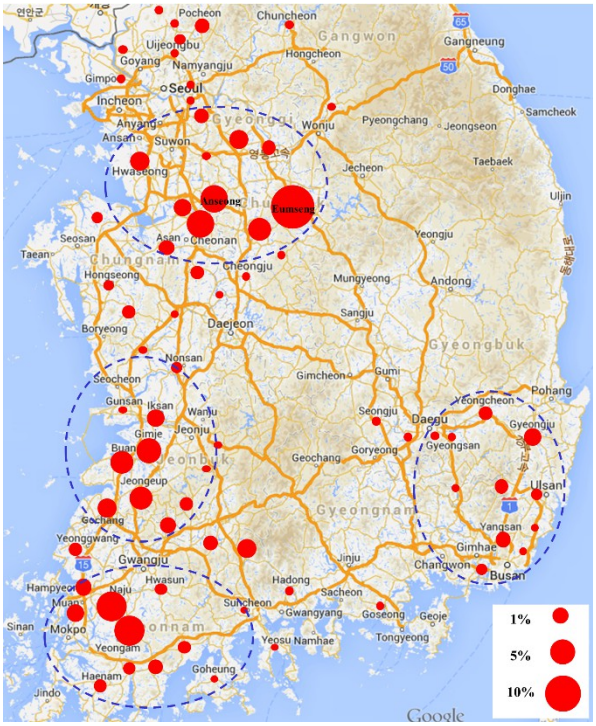

Figure 2: HPAI Outbreak Status Across the Country

TID	EID	Items
1 (03/04)	1	JeanNam.Yeongam
1 (03/04)	2	Jeonbuk.Iksan
1 (03/04)	3	Gyeonggi.Pyengtak
2 (06/07)	1	JeanNam.Naju
2 (06/07)	2	Jeonbuk.Iksan
2 (06/07)	3	Gyeonggi.Pyengtak
3 (08)	1	Jeonbuk.Kimje
3 (08)	2	Jeonbuk.Iksan
3 (08)	3	Gyeonggi.Anseng
3 (08)	4	JeanNam.Yeongam
4 (10/11)	1	JeanNam.Naju
4 (10/11)	2	Jeonbuk.Iksan
4 (10/11)	3	JeanNam.Yeongam
5 (14/15)	1	Gyeongbuk.Yangsan
5 (14/15)	2	Gyeonggi.Anseng
5 (14/15)	3	JeanNam.Yeongam
5 (14/15)	4	JeanNam.Naju
5 (14/15)	5	Jeonbuk.Iksan

Table 2: Example Data set for Sequential Pattern Mining

3. Results

In this section, we test the feasibility and applicability of the proposed system by describing the experimental results that were applied to the HPAI outbreak data using R tools [8]. We are interested in identifying the sequential dissemination routes of HPAI outbreaks in South Korea. From our trial experiments, we describe a process based on a certain scenario that elicits information that generates the sequential dissemination routes of HPAI outbreaks from "Gochang" within a specific period. We used the HPAI occurrence data set in 2014 with the "arulesSequences" package in R tools for sequential pattern mining analysis. Table 3 presents sample of the obtained sequential pattern mining rules (mininum support $\geq$ 0.2) without considering the nearby sites (see Figure 2) to "Gochang". Concerning the incubation period [1,9] , we can observe that the outbreak is propagated rapidly over a wide area during the first

two weeks and is spread sparsely after two weeks; it is propagated throughout the nation within six weeks [9]. Therefore, we can state that it is critical to initiate preventive actions against HPAI in the first two weeks. Finally, we depicted the sequential propagation routes of HPAI outbreaks from "Gochang" in a GIS version of South Korea (see Figure 3).

Index	Sequential routes of HPAI outbreaks
1	Jeonbuk.Gochang(14.01.16) → Jeonbuk.Buan(14.01.17) → Chungnam.Buyeo(14.01.25)
2	Jeonbuk.Gochang(14.01.16) → Jeonbuk.Jeongeup(14.01.24)
3	Jeonbuk.Gochang(14.01.16) → Jeonnam.Haenam(14.01.26)
4	Jeonbuk.Gochang(14.01.16) → Jeonnam.Naju(14.01.28) → Jeonnam.Yeongam(14.01.30)
5	Chungnam.Buyeo(14.01.25)→Chungnam.Chenan(14.01.28)→Chungnam.Chengyang(14.02.17)
6	Chungnam.Chenan(14.01.28)→Chungbuk.Jinchen (14.01.29) | Gyeonggi-do.Hwaseong(14.01.30) | Gyeonggi-do.Anseng(14.02.13)→Gyeonggi-do.Pyeongtak(14.02.25)
7	Jeonbuk.Jeongeup(14.01.24) → Jeonbuk.Imsil(14.01.29) → Gyeongnam.Miryang(14.01.30)
8	Chungbuk.Jinchen(14.01.29) → Chungbuk.Emseong(14.02.07)
9	Jeonbuk.Jeongeup(14.02.17) → Jeonbuk.Gimje(14.02.19) → Chungnam.Nonshan(14.02.24)
10	Jeonnam.Yeongam(14.02.17)→Jeonnam.HamPyeong(14.02.24)→Jeonnam.Yeonggwang (14.02.28)

Table 3: Example of Obtained SequentialPattern Rules

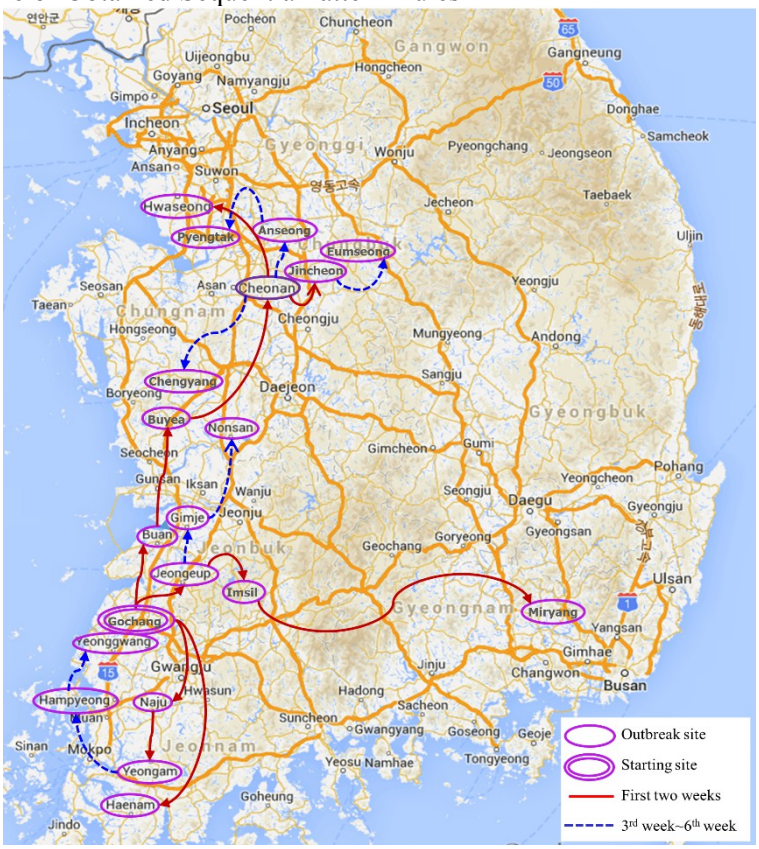

Figure 3: Sequential routes of HPAI outbreaks from "Gochang" in GIS.

4. Discussions and Conclusion

The detailed analysis of HPAI occurrences and early prediction of the HPAI outbreak sequence are important issues in the livestock industry. In this paper, we proposed a new design methodology for the analysis of HPAI in South Korea. The core of the methodology is the construction of an integrated data analysis model using HPAI occurrence data accumulated in a data warehouse over an extended period. The proposed analysis model elicited useful information that generated the sequential dissemination routes of HPAI outbreaks by applying sequential pattern mining. We tested the feasibility and applicability of the proposed HPAI

analysis model by implementing an HPAI analysis system using R tools and then applying it to the analysis of HPAI outbreaks in South Korea.

To the best of our knowledge, this is the first report of OLAP analysis for accumulated long-term livestock disease occurrence data and data mining analysis for the dissemination routes analysis of disease outbreaks. The application of big data analytics including data mining methods is appropriate considering the continuous and large incoming data stream that is characteristic of the new era of big data. Further testing and refinement of the proposed system is required and is a part of our ongoing research. We also plan to increase the power of our prototype system by embedding various data-mining analysis techniques in it.

References

[1] H.-R. Kim, Y.-K. Kwon, I. Jang, Y.-J. Lee, H.-M. Kang, K.-H. Lee, et al. *Pathologic changes in wild birds infected with highly pathogenic avian influenza a (H5N8) viruses, South Korea, 2014* [J]. Emerging infectious diseases. 21(5): 775-780(2015)

[2] I.-H. Seo, I.-B. Lee, O.-K. Moon, N.-S. Jung, H.-J. Lee, S.-W. Hong, et al. *Prediction of the spread of highly pathogenic avian influenza using a multifactor network: Part 1–development and application of computational fluid dynamics simulations of airborne dispersion* [J]. Biosystems engineering. 121(1): 160-176(2014)

[3] H.-J. Lee, K. Suh, N.-S. Jung, I.-B. Lee, I.-H. Seo, O.-K. Moon. *Prediction of the spread of highly pathogenic avian influenza using a multifactor network: Part 2–comprehensive network analysis with direct/indirect infection route* [J]. Biosystems engineering. 118(1): 115-127(2014)

[4] Tuncer, N., and Le, T. *Effect of air travel on the spread of an avian influenza pandemic to the United States* [J]. International journal of critical infrastructure protection. 7(1): 27-47(2014)

[5] KAHIS (Korean Animal Health Integration System). *http://www.kahis.go.kr/home/lkntscrinfo/selectLkntsOccrrncList.do* Accessed December 10, 2015

[6] J. Han, M. Kamber, and J. Pei. 3rd ed., *Data Mining: Concepts and Techniques* [M]. Morgan Kaufmann Publishers, MA Burlington. 125-148,243-272(2012)

[7] M. J. Zaki. *SPADE: An efficient algorithm for mining frequent sequences* [J]. Machine learning. 42(1): 31-60(2001)

[8] R arulesSequences. *https://cran.r-project.org/web/packages/arulesSequences/index.html A*ccessed December 10, 2015

[9] H. Yoon, O.-K. Moon, W. Jeong, J. Choi, Y.-M. Kang, H.-Y. Ahn, et al. *H5N8 highly pathogenic avian influenza in the republic of Korea: Epidemiology during the first wave, from January through July 2014* [J]. Osong public health and research perspectives. 6(2): 106-111(2015)