PKGAWAS: A Knowledge Services and Allergy Early Warning System of Pollinosis Based on Cross-Border Data Integration

Xiaolei Xiu, Sizhu Wu, Jiawei Cui, Xiaokang Sun, Qing Qian*
Institute of Medical Information
Chinese Academy of Medical Sciences & Peking Union Medical College
Beijing, China
e-mail: xiu_xiaolei@163.com; wu.sizhu@imicams.ac.cn; cui.jiawei@imicams.ac.cn;
sun.xiaokang@imicams.ac.cn; qian.qing@imicams.ac.cn

Abstract— **In order to meet China's growing demand for knowledge services and allergy early warning of pollinosis, the paper initially explores and develops a system, called "Pollinosis Knowledge Graph and Allergy Warning Analysis System" (PKGAWAS). The system's data comes mainly from professional websites, Chinese Wikipedia and texts, which cover four fields: medicine, agriculture, forestry, and geography. In order to effectively implement multi-source data integration, this paper first uses machine and manual methods to collect data from literature, books, and websites. Then, we store the collected data in a temporary database. After the data is normalized, we store data in different categories, and then integrate the data based on the relationships between entities. Specifically, the paper uses property graph for knowledge representation of the knowledge graph, and other data are analyzed from the three dimensions of space, time, and disease. PKGAWAS not only can provide users with a full range of knowledge services and help, but also has important physical and practical significance for promoting "Healthy China 2030".**

Keywords-PKGAWAS; open data; data integration; knowledge graph; allergy early warning.

I. INTRODUCTION

Pollen allergy first appeared in the 19th century. In 1828, Bostock published its first report on hay fever [1]. Subsequently, in 1873, Blakely proved that pollinosis was caused by pollen from grasses [2]. In the second half of the 20th century, the prevalence of allergic respiratory diseases, such as allergic asthma and allergic rhinitis increased dramatically, affecting millions of people [3]. Pollen allergy is a common disease of allergies, and the incidence rate has increased year by year, seriously affecting human health.

Pollinosis has become a veritable epidemic. It is estimated that due to the increase in urban green areas, more than 50% of the population in industrialized countries will experience hypersensitivity in the next 20 years [4], but it is difficult to cure pollinosis.

In order to effectively reduce the incidence of hay fever, we need to do a good job in prevention [5]. Airborne pollen is routinely monitored in many parts of the world, such as North America and Europe, and the first limited network has also been created for monitoring airborne allergen concentrations [6]. In contrast，the pollen allergy situation in China is grim, but there is no specialized pollinosis knowledge service and allergy early warning website. In China's 18-64 year-old population, the incidence of pollen allergy is 0.5 ~ 1%, and high-risk groups may even reach 5% [7]. However, China only conducted a nationwide survey on the distribution of airborne allergenic pollen in the 1980s. Nearly 30 years later, historical data cannot accurately analyze and predict future pollen concentrations.

But allergic diseases are now receiving the attention of the national government. On October 25, 2016, the CPC Central Committee and the State Council issued and implemented the "Health China 2030" Plan. The purpose of this plan is to promote the construction of a healthy China and improve people's health. During the same year, Beijing Smart Park Summit, Gao Wei, deputy director of the Beijing Municipal Bureau of Landscaping, announced that during the "13th Five-Year Plan" period, Beijing will launch the construction of smart garden system to provide citizens with personalized recommendation services, such as allergens Early warning and others.

In order to meet the growing demand for knowledge services and early warning of pollinosis, this paper attempts to design and develop a system, called Pollinosis Knowledge Graph and Allergy Warning Analysis System (PKGAWAS). The study uses the methods of data integration to secondary develop and utilize of Web data and texts, which span four fields of medicine, meteorology, forestry and geography. Additionally, PKGAWAS can provide users with scientific knowledge of pollen allergy, knowledge graph and allergy early warning, personalized customization services and so on. The development of PKGAWAS is of great theoretical and practical significance for the pollen allergy knowledge service and allergy early warning website construction in China.

The rest of the paper is organized as follows: Section II describes the system architecture. Section III introduces the system data management process, including data source, data collection and preparation, data integration. Section IV presents the system interface. Conclusion and the aspects for future works are provided in Section V.

II. METHODOLOGY

A. Architecture

As it has been mentioned in Section I, the main goal of the PKGAWAS-based on Cross-Border Data Integration is to meet the growing demand for knowledge services and

 ISBN: 978-1-61208-631-6

allergy early warning of pollinosis. According to this demand, the system needed to fulfill the following main functional requirements:

1) Cross-border data integration: System data come from the fields of medicine, geography, agriculture and history.

2) Multidimensional correlation analysis: It aims to dig deep into the relationship between pollen concentration and space, climate, time and the impact of allergens and regional differences on hay fever.

3) Knowledge organization: Construct a disease-centered dynamic interactive knowledge graph.

4) Allergy early warning.

5) Dynamic interactive visualization.

6) Personalized custom service.

Based on the above functional requirements, the overview of the system architecture of the PKGAWAS are shown in Figure 1. The system mainly has four layers: support layer, data storage layer, functional layer, application layer.

B. System Design

In order to make the website simple and intuitive, the site link level cannot exceed three. The overall structure of the system is shown in Figure 2. PKGAWAS has a total of three layers.

- The first layer is the home page. On the front page, we can see the main functions of PKGAWAS at a glance. In addition, users can not only search for pollen, doctors, hospitals, medicines and pollen allergy related diseases, but also direct access to doctor's database, hospital database, medicine database and Pollen database.
- The second layer is the columns page. This system has five columns page, which are pollen allergy related diseases, knowledge graph, airborne allergenic pollen map, allergy early warning, and about. Pollen allergy related diseases include six diseases, such as pollinosis, allergic rhinitis, bronchial asthma. The system's knowledge graph is dynamically interactive, besides it also has intelligent statistics and related recommended functions. Airborne allergenic pollen map introduces regional, monthly and disease spectrum of pollenfrom a national and local point of view. In the allergy early warning page, PKGAWAS provides users with allergy tracker, future forecast, literature allergy prediction and other services.
- The third layer is the content page.

III. DATA MANAGEMENT

A. Data Source

The data of this study mainly comes from professional websites, Chinese Wikipedia and texts, which cover four fields: medicine, agriculture, forestry, and geography. Professional websites include Clinical Medicine Knowledge Base (CMKB) [8], Beijing Meteorological Service [9], China

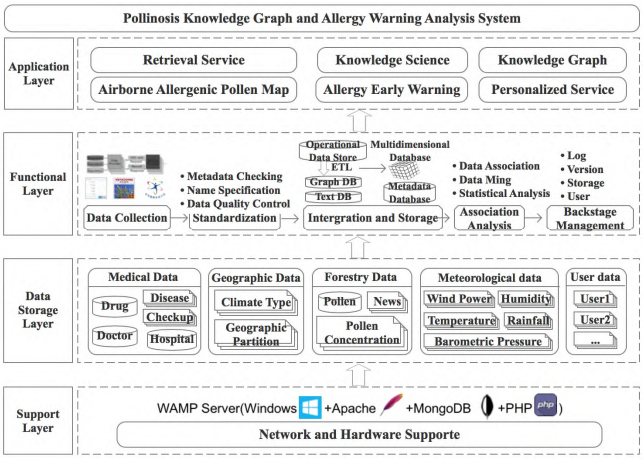

Figure 1. Overview of the system architecture

Weather Network [10], China Food and Drug Administration (CFDA)[11], and Chealth online [12] and so on. The text data refers to literature and a book entitled "Color Atlas of Air-borne Pollens and Plants in China". For literature data, we searched the literature related to pollen allergy from 2000 to 2017 from Wanfang data and the National Knowledge Base Database (CNKI). Chinese Wikipedia refers to Baidu baike and Hudong baike.

B. Data Collection and Preparation

In order to accurately and comprehensively collect data, we must first identify which data to collect and plan the representation of knowledge.

1) Preliminary: The paper adopts the property graph model to perform the knowledge representation of the knowledge graph. The system's property graph is constructed manually and then evaluated by experts. It contains two parts: nodes and relationships, such as Figure 3.

a) Node: Nodes s are the entities in the graph. They can hold any number of attributes (key-value-pairs) called properties. Nodes can be tagged with labels representing their different roles in your domain. In addition to contextualizing node and relationship properties, labels may also serve to attach metadata—index or constraint information—to certain nodes.

b) Relationship: Relationships provide directed, named, semantically relevant connections between two node-entities. A relationship always has a direction, a type, a start node, and an end node. Like nodes, relationships can also have properties [13].

The conceptual layers of the knowledge graph of this system include: diseases, complications, doctors, hospitals, drugs, medical examination methods, and drug companies. For the entities in concepts and concepts in the knowledge graph, the paper uses a top-down and bottom-up approach to construct the knowledge graph. However, the top-down approach is not constructed by building the top relational ontology, but directly by the property and entity-to-entity relationship in the property graph.

2) Data collection： Data collection refers to the process

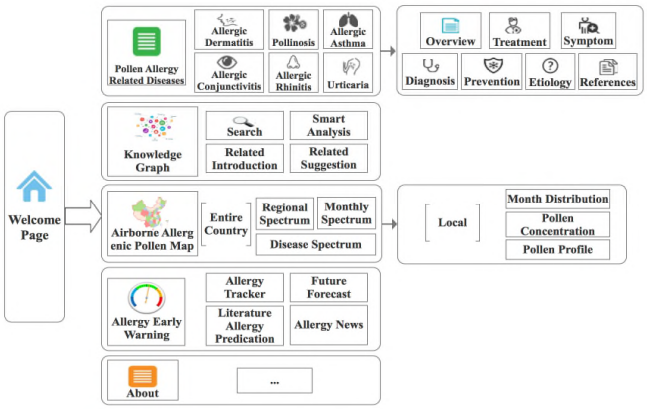

Figure 2. Overall structure of the system

of identifying, selecting, and collecting data from data sources. The data collection method of this study is mainly machine collection, and manual collection of text data is auxiliary.

c) Web Crawler: Using Web crawler technology to crawl entities, property and entity-to-entity relationships from websites, such as the CFDA, Chealth online, CMKB and Chinese Wikipedia and so on. The data obtained using Web crawling techniques is mainly medical data and geographical data.

d) Data interface: We uses the API approach to collect pollen concentration, weather information, such as temperature, humidity, barometric pressure and wind power at the Beijing Meteorological Bureau and China Weather Network.

e) Manual collection: The paper uses manual extraction to extract data from literature and books. This is because the amount of data in the literature is small, and the content required is cluttered. Manual extraction can improve accuracy. Through manual extraction, we collected pollen-related information and medical data.

C. Data Integration

This brief data integration process is shown in Figure 4. First, we store the collected data in a temporary database. After data cleansing, conversion, and other normalization processes, we integrate and store the data. In this study, we store data in different categories. The conceptual data and entity data used to construct knowledge graphs are stored and integrated by Neo4j graph database [13]. Then, other pieces of information are stored in a relational database.

The paper integrates the data from three aspects, such as the relationship between entities, different dimensions and application integration.

- The relationship between entities. The paper has constructed a knowledge graph that uses the relationships between entities. Knowledge graph would dynamically present and manage the data that cannot be statically stored and displayed, allowing users to search for the medical information on pollinosis and conduct data mining. There are semantic relationships between diseases, doctors, hospitals, medicines, checkups and pharmaceutical companies, such as “belong to”, “treat”, “examine”. The paper first constructs a property graph for knowledge graph to quickly and accurately construct a disease-centered knowledge graph.

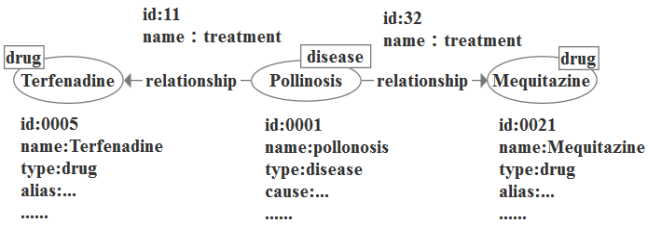

Figure 3. Data Model of Property Graph

- Different dimensions. We analyzed factors affecting pollen concentration from different dimensions, such as space, time and disease. First, the paper conducts a statistical analysis of pollen concentration and geographic partition, climate types, months, related diseases, allergens and other data. Then, an airborne allergenic pollen map including the regional spectrum, the monthly spectrum, and the disease spectrum was constructed. In addition, we displayed in a visual form, such as a nightingale's rose diagram, a doughnut chart, map and so on.
- Application integration. In order to make every page of this system not isolated, this study integrates applications so that they are "alive" and can be associated with other Web pages.

IV. AIRBORNE POLLEN PREDICTION MODEL

There are few studies on airborne pollen prediction model. Several popular algorithm models include neural network models [14][15] and multiple regression algorithms [16][17]. Considering many factors, such as data volume, localization, and authoritativeness, this system adopts the airborne pollen prediction model announced by the Tianjin Meteorological Bureau, which is funded by the China Maritime Affairs Bureau's new technology promotion project “pollen detection and service”.

This is a staged prediction model. According to the high, low and stable development trend of pollen concentration, the whole pollen period is divided into 6 stages: In the first stage, the pollen begins to peak in spring; The second stage is the peak period to the spring sub-peak; The third stage the sub-peak to June; From the middle of June to the beginning of August, it is the fourth stage; The fifth stage is from the late August to peak in autumn; after the peak period, when the pollen is over it drops to sixth stage. The multiple regression model is as follows:

$$\Psi_1 = 8565.13 - 0.33H_1 + 0.12H_2 - 0.25H_3 + 3.18T_\alpha^2 - 41.53T_\alpha + 58.77P_3 + 44.61\ T_{min10} - 25.35\ T_{max10} + 17.37V_{\alpha 10} - 8.09P_{10} \quad (1)$$

$$\Psi_2 = -397.84 + 0.14H_1 + 0.13H_2 - 0.16H_3 + 19.716V_{max} + 0.39T_{min8} - 11.29T_{max10} + 5.82T^2_{\alpha 10} - 197.62T_{\alpha 10} + 2.23P_{10} \quad (2)$$

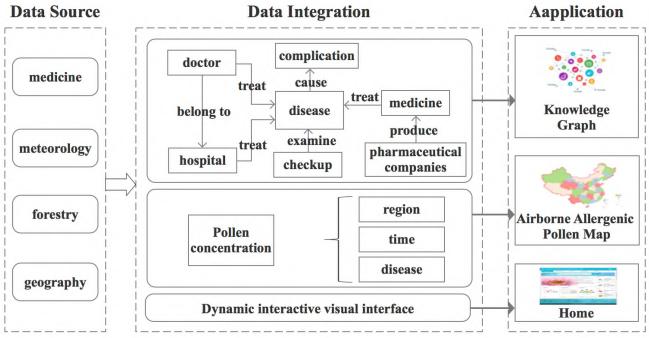

Figure 4. System Data Integration Flow Chart

$$\Psi_3 = 4181.61 - 0.01H_3 - 3.65\Pi + 2.98T_{max3} + 9.35P_8 + 2.84T^2_{min10} - 65.55T_{min10} - 7.1668T_{\alpha 10} - 14.22V_{max10} \quad (3)$$

$$\Psi_4 = -361.40 + 0.06H_2 + 1.67T_{\alpha} - 0.24f + 0.77T^2_{min10} - 33.72T_{min10} - 0.03T_{max10} - 0.353T_{\alpha 10} + 0.75P_{10} \quad (4)$$

$$\Psi_5 = -4231.271 - 0.337H_1 + 0.137H_2 + 0.39H_3 + 10.26V_{max} - 0.61T^2_{\alpha 10} + 32.98T_{\alpha 10} + 0.63P_{10} + 9.84V_{a10} + 3.7292P_{10} \quad (5)$$

$$\Psi_6 = 1183.49 - 0.14H_1 + 0.10H_3 + 0.63T^2_{\alpha 10} - 16.1851T_{\alpha} - 1.04P_{10} - 0.45f_{10} \quad (6)$$

where Ψ1, Ψ2, ···, Ψ6 are predicted values of pollen concentration. H represents pollen concentration. T_a is average temperature, and T_{min} represents minimum temperature. Maximum temperature is represented by T_{max}. R represents precipitation and P is average pressure. Average relative humidity is represented by f and average wind speed is represented by V_a. V_{max} represents maximum wind speed. Besides, the digital subscript indicates the number of days before the forecast, and T_{min10} indicates the average minimum temperature in the first 10 days of the forecast. If there is no data subscript, it means the next 72 h variable.

V. USE CASE

Our system is implemented and publicly accessible [18]. The home page of the website is shown in Figure 5.

Figure 5. Welcome page of the PKGAWAS

VI. CONCLUSION

Compared with other existing websites, PKGAWAS has the following four aspects of innovation: a) multi-source cross-border data integration; b) multi-dimensional data association analysis; c) dynamic interactive knowledge graph; d) personalized custom service. However, due to the lack of pollen concentration data, there is still much work to be done in the future. First, we will use a crowdsourcing approach to collect pollen concentrations in the country. The second is to seek cooperation from the China meteorological administration to jointly carry out early warning service for pollen allergy.

ACKNOWLEDGMENT

The authors thank Chealth Online for providing medical data. The work of the authors is supported by Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College.

REFERENCES

[1] J. Bostock, “Case of a Periodical Affection of the Eyes and Chest, ” Annals of Allergy, vol. 18, 1960, pp. 894.

[2] C. H. Blagkley, “Experimental researches on the causes and nature of catarrhus aestivus (hay-fever or hay-asthma),” Blackley, 1959, pp.57.

[3] G. D. Amato and L. Cecchi, “Effects of climate change on environmental factors in respira- tory allergic diseases,” Clin Exp Allergy, vol. 38, 2008, pp. 1264–1274.

[4] X. D. Xiao, “Talk about flowers no longer change color,” Captial Medicine, vol. 9, 2017, pp. 60–62.

[5] L. P. Dai and C. Lu, “The Pollen and Its Measurement Technique in Spring,” Meteorological Monthly, vol. 12, 2000, pp. 49–52.

[6] M. Smith, U. Berger, H. Behrendt and K. C. Bergmann, “Pollen and pollinosis,” Chem Immunol Allergy, vol. 100, 2014, pp. 228-233.

[7] Q. Y. Wei, “Diagnosis and Treatment of Pollinosis,” Chinese Journal of Practical Internal Medicine, vol. 32, 2015, pp. 89–91.

[8] Institute of Medical Information, Chinese Academy of Medical Sciences, “Clinical Medicine Knowledge Base,” 2014, http://www.cmkb.cn.

[9] Beijing Meteorological Bureau, “Beijing Meteorological Service,” 2008, http://www.bjmb.gov.cn.

[10] CMA Public Meteorological Service Centre, “China Weather Network,” 2008, http://www.weather.com.cn.

[11] China Food and Drug Administration, “China Food and Drug Administration,” 2013, http://app1.sfda.gov.cn/datasearch/face3/dir.html.

[12] Institute of Medical Information, Chinese Academy of Medical Sciences, “Chealth”, 2014, http://www.chealth.org.cn.

[13] Neo4j, “What is a Graph Database? ” 2017, https://neo4j.com/developer/graph-database/.

[14] J. A. Sánchez-Mesa, C. Galan, J. A. Martínezheras and C. Hervásmartínez, “The use of a neural network to forecast daily grass pollen concentration in a Mediterranean region: the southern part of the Iberian Peninsula,” Clinical & Experimental Allergy, vol. 32, 2002, pp. 1606–1612.

 ISBN: 978-1-61208-631-6

[15] M. Puc, “Artificial neural network model of the relationship between Betula, pollen and meteorological factors in Szczecin (Poland),” International Journal of Biometeorology, vol. 56, 2012, pp. 395-401.

[16] K. R. Kim et al., “A biology-driven receptor model for daily pollen allergy risk in Korea based on Weibull probability density function,” International Journal of Biometeorology, vol. 61, 2016, pp. 1-14.

[17] Z. L. Wu, et at., “Study of Airborne Pollen Prediction Model,” Meteorological Science and Technology, vol. 35, 2007, pp. 832-836.

[18] Institute of Medical Information, Chinese Academy of Medical Sciences, “Pollinosis Knowledge Graph and Allergy Warning Analysis System,”2017, http://114.255.123.93:6606/Pollen/.

 ISBN: 978-1-61208-631-6