Construction of Machine News Production System Based on Graph Database and Natural Language Generating Technology

Haiyun Han[1,2]
1. Communication Strategy of Co-innovation Center
Communication University of China
2. Fire Engineering Department
Chinese People's Armed Police Force Academy
Beijing, China
Email: 353614438@qq.com

Pengzhou Zhang
School of Computer
Communication University of China
Beijing, China
Email: zhangpengzhou@cuc.edu.cn

***Abstract*—The using of machine to write news automatically has greatly increased news' output and efficiency in media organizations. However, the number of news subject applicable to machine news writing technology is limited, and the form of machine news expression is relatively single. By analyzing the writing methods and principles of machine news and the characteristics of news materials storage database, as well as the types, sources and usage requirements of news information data, it was proposed that the data storage methods and production of machine news can be improved by using graph database tools and natural language generation technology; and then a machine news production system composed of four levels and nine modules was constructed, and the system's design ideas, logical framework and operation process were elaborated. The development and application of that system will broaden the field of machine news production and enrich the presentation of machine news.**

Keywords- machine news; graph database; natural language generation; system framework

I. INTRODUCTION

A. Background

In order to increase the production and efficiency of news and to meet diversified and personalized needs of the long tail users, "machine news" has emerged at the right moment and has become one of the representative applications of artificial intelligence in the media industry. As early as 2010, Forbes website, Associated Press, Bloomberg, Los Angeles Times and other foreign media have implemented "robots writing news" in sports, finance, and weather fields [1,2]. Since 2015, Tencent and Xinhua News Agency have successively taken the lead in Chinese media to experiment on news writing robots named "Dream writer" and "Quick writing". In 2016, the writing robots developed by China's headline laboratories generated more than 200 news and briefs just six days before the start of the Rio Olympic Games [3]. It can be seen that the news automatic writing robots has an obvious advantage on writing speed and massive production than the manual editing, and then, reporters can be free from the work of single, repetitive fact reporting, and save more energy to product high-quality and in-depth news.

B. Problem presentation

The application of machine news writing has been expanded from original sports events to financial news and financial reports, and to weather forecast, disaster news, crime news and so on. At the same time, the sources and data form of machine news material are diversity, and the relationship among entities in news theme is becoming increasingly complex. Therefore, it is the research forefront to solve how to store, query, and efficiently invoke those rich data resources and automatically manufacture variety of news works on rich themes. This paper proposed to use of graph database theory and techniques to store and retrieve news data from multiple sources, and to construct a machine news writing system with a wider applicable fields and richer news representations based on natural language generation technology. [4].

II. RELATED RESEARCH

A. Methods on machine news writing

There are two basic ways of machine news writing. One is filling text or template-filling writing. The other is building natural language generation system with data mining and computational linguistics methods and techniques to automatically generate understandable text by algorithms [5].

1) Machine news writing with templates

News writing with template is a relatively mature technique of machine news production, and it is widely used. For example, the Stats Monkey software was originally developed by Prof. Kris Hammond and his students at the Northwestern University's Intelligent Information Laboratory, which was used for reporting on baseball games; And the machine writing platform QUILL was established by Narrative Science, which specialized in writing financial news reports and providing financial data analysis reports for companies; as well as Wordsmith platform which was launched by Automated Insights [6]. These writing software and systems are all used with template to write, and the basic principles of machine news writing with template can be divided into 5 steps, as shown in Fig.1.

978-1-5386-5500-9/18/$31.00 ©2018 IEEE
DOI 10.1109/ICISCE.2018.00045

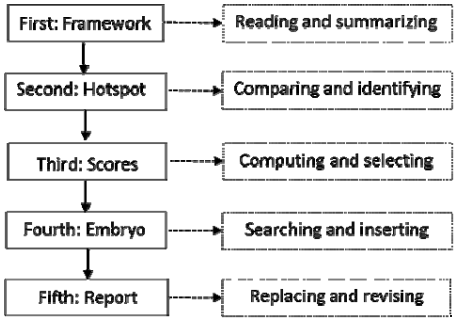

Figure 1. The principle of machine news writing

2) Machine news writing based on natural language generation technology

The machine news writing system based on natural language generation technology is a machine news production method that is actively being explored [7]. Natural language generation (NLG) is a branch of artificial intelligence and computational linguistics. NLG is a computer model based on language information processing, which selects the abstract concepts and implement some semantic and grammatical rules to generate text [8-10]. The basic workflow of NLG system is shown in Figure 2. It usually includes content planning, micro-planning, and surface generation [11, 12]. The content plan includes two tasks: content selection and structure construction. Specifically selecting the information data needed to generate text from the mass information, and then determining appropriate text structure for semantic representation and content arrangement. Micro-planning is the local detail planning of texts, including Lexicalization, Aggregation, and Referring Expression Generation. Surface generation mainly involves structure implementation and language implementation, the task is to map text description from micro-planned to the surface text composed of written words, punctuation, and structure annotation information [13-14].

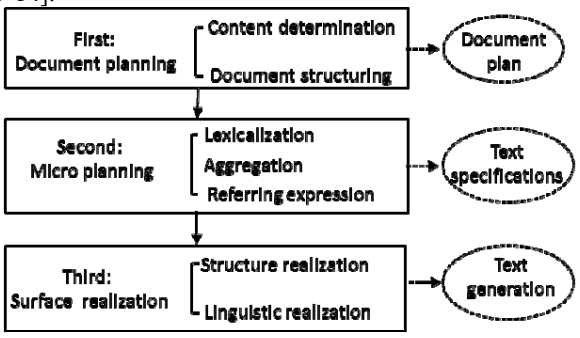

Figure 2. The basic workflow of NLG system

B. Database for storing news information

The information of the news material is stored in the database for searching and transferring by the machine writing system. A database is a warehouse that organizes, stores, and manages data in accordance with the logical structure of data. It is called a data model that reflects and implements the logical structure of data. At present, there are three more popular data models: the relational structure model established by relational theory, the hierarchical structure model and mesh structure model established by the theory of graph theory, thus forming two types of relational databases and non-relational databases [15].

1) Relational Database

Relational database based on a relational model which processes data with mathematical concepts and methods of collection algebra. Relational database is usually composed of a number of two-dimensional row list grids that can be linked to each other. The popular relational databases include Oracle, DB2, PostgreSQL, Microsoft SQL Server, Microsoft Access, MySQL, and Inspur K-DB. Currently. In recent, most of template-based machine news writing systems apply relational database to directly search and reference data from two-dimensional tables to fill in a writing template. This kind of database has good performance for news content with simple data relations, such as scores of the game, value of stock, and data parameters such as temperature and humidity in the weather forecast.

2) Graph Database

The graph database is a non-relational database which applies graph theory to store the relationship information between entities. The structure of graph database consists of a series of nodes and directed connections that reflect the relationships between the nodes, as shown in Figure 3. Each node in the graph represents an entity and records its attribute information. The node connection in the figure represents the relationship [16]. Frequently-used graph databases include Neo4j, AllegroGrap, GraphDB, InfiniteGraph, OrientDB, InfoGrid, and HypergraphDB. Graph databases can be used to store complex knowledge bases, social relationships, etc. [17]. News information data can be stored in a graph database by converting news information into entities and entities.

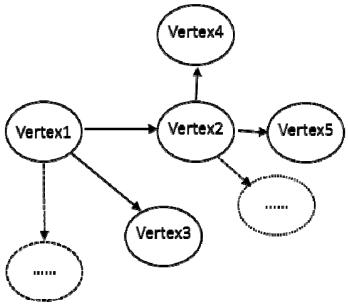

Figure 3. Data structure model of graph database

3) Advantages of Graph Databases Compared to Relational Databases

Compared to relational databases, graph databases have two main advantages for news data management:

a) Good at storing news data with complex relationship

Because of two-dimensional characteristic of relational database, a series of tables would be created to record multi-to- many relationships of different entities, and these association tables often do not record information. If a piece of news including more entity objects and these relationships are complex, it is necessary to construct a large number of two-dimensional tables and create multiple associated tables among them. However, in view of the above problems, graph database requires only connection links to indicate that there are different relationships among two entities [19]. In other words, the various association tables in relational database can be replaced by relations containing attributes in graph database, therefore graph database can realize more abundant data relation description and calculation with more intuitive and convenient format.

b) Efficient data retrieval and query

When relational database is used to characterize complex relationships of entities in news topics, it need to construct a number of two-dimensional tables and associated tables, and there may be data crossover and repetition. When querying and computing data, a large amount of computing resources will be consumed in many forms, and the computation efficiency is low. by contraries, Graph database representation entity relationship is concise and clear, and the calculation method of data is intuitive, and it has the advantages of convenient and fast data query [20, 21].

III. CONSTRUCTION OF MACHINE NEWS PRODUCTION SYSTEM

The machine news production system was built with web data mining, intelligent semantic analysis, graph database and natural language generation and other technologies to collects and processes news information data and automatically produces news content.

A. Basic ideas

First, to collect domain-specific knowledge and news material with data mining technology and intelligent semantic analysis technology by human or algorithms. Then, to establish knowledge concept models and define these as computer-recognized entity and relationship, and to establish mesh data models and store these into graph databases which are capable to automatically search for information and extracts information; Next, to provide two methods of machine news writing, which are news writing templates and NLG for news writing. Wherein, the method news writing template only needs to retrieve corresponding data from the database, and another news writing method based on NLG needs to complete automatically news content planning according to the types of users' requirement, such as fact description, reason investigation, comparative review, etc. Finally, the news prototypes will be further improved with adding or concatenating linguistic corpus of different language styles after analysis of user types, as well as matching richer pictures or multimedia, so as to complete the final presentation of the press release.

B. System Framework

The logical framework of the machine news production system is shown in Fig. 4. This system consists of four levels and nine modules: The first layer is information collection of news writing, which includes acquisition module of relatively fixed knowledge in specific fields and collection module of dynamic news theme information; The second layer is news information graph database, including the news theme knowledge ontology database, news dynamics database and user model database; the third layer is news content production, including news template filling module and NLG news text module. The fourth layer is typesetting and distribution for news works, and it is divided into manual checkout module and personalized custom distribution module.

C. Workflow

The working process of the machine news production system includes basic steps: data collection and cleaning, data storage, knowledge base construction, content planning, text generation, multimedia manuscript generation, and multi-platform news production and distribution.

1) Data acquisition and processing

Before news data collection, it is important to determine data range required for news writing firstly, and to make clear which data is relatively stable or even steadiness for long-term, and which data is dynamically changing, those will decide whether to crawl data with once or periodically. The collection of knowledge information for a specific news field can be combined by manual collection and computer automatic acquisition. The collection of dynamic information data for news can be automatically collected by computer, and the time node or cycle of data acquisition is also analyzed. After collecting data from all sources, it is necessary to perform simple data cleaning, remove redundant information, and ensure the integrity of the data.

2) Building graph databases

Effective data acquired after data acquisition and cleaning are required to be stored for using. The dynamically updated information data can be stored in relational databases or graph databases. The relatively invariable knowledge data will be transferred into knowledge model library which can achieve four core contents with knowledge graph technology: knowledge extraction, knowledge representation, knowledge fusion and knowledge reasoning. The factual knowledge acquired from the above data collection and the dynamic information obtained from data mining will become an important source of machine news writing materials.

3) Content selection and structure planning

For machine news writing based on template filling, different topics and contents correspond to different templates, and the text structure has been planned in advance. For machine news writing based on natural language generation technology, computers will select content and structure with algorithm in real time according to news topics. When users input instructions of news topics and news types, computers will find and filter the data of entities

and relationships in the related database by semantic analysis and text classification algorithm.

4) Text Generation

In text generation stage, the main task of machine news writing based on template is to fill appropriate phrases in the blank space. In order to make the filled phrases more professional and readable, it is necessary to set specific rules on generation phrase in different news theme. After the phrases are generated, the complete text content can be generated in the previously selected template. The text generation based on natural language generation system includes two steps: Micro planning and surface realization. Vocabularies should be selected and sorted in the part of micro planning, then passed to the part of surface realization for text generation.

5) Multimedia manuscript generation

After completing the main stage of machine news text writing, it is possible to enter the automatic generation phase of multimedia manuscripts. In order to make the news automatically generated by the machine more readable and interesting, besides the text, multimedia data such as pictures and videos related to news should also be added. For the results of data analysis and mining in the writing process, data visualization can also be used to present the user, and at the same time, user interaction is enabled, so that the machine automatically generates more diversified news content.

6) News manuscripts typesetting and personalization for users

The presentation of news manuscripts in different forms of media communication should also be different, such as newspapers, television, radio and internet, and mobile phone each have their own performance characteristics. It is necessary to manually edit, check and layout adjustments for machine news content according to the target media. The use of WeChat, public number, news client, micro blogging and other channels for distribution and dissemination news in the new media communication environment, which requires machine news style should also be personalized according to the user model library including interest tags, geographic location, reading style information. Accurate push makes the news generated by machine writing to maximize its value, and at the same time, it can also create a good reading experience for users.

IV. CONCLUSION

Grape database technology provides an effective tool for news information management with efficient and convenient. The natural language generation technology lays a theoretical foundation for real machine news writing. The machine news production system constructed based above technologies and theories have the following features:

- Knowledge and dynamic news data as well as user characteristics and relational model are stored by graph database which can better represent the complex relationships between entities, and at the same time, it is well-suited for machine news writing to retrieve data and analysis relationship among data.
- The machine news writing system based on natural language generation technology including content planning, micro-planning and surface realization will break through the limitations of news topics or fields which are involved by original machine news with production methods of template-filled and expand news types from single fact statement news to comparative evaluation news and reason explanation news.
- Customized machine news according to user's characteristics will be produced with the information database of users and their relations; as well as machine news can also display diversified works by loading multimedia and pictures.

ACKNOWLEDGMENT

This research is supported by the project "Research and demonstration application of Chinese News automatic writing system based on knowledge map of brain". The financial support of Beijing Science and Technology Commission is appreciated.

REFERENCES

[1] Clerwall C. "Enter the Robot Journalist: Users' perceptions of automated content" , Journalism Practice, 2014 (ahead-of-print): pp. 1-13.

[2] Bai L. "The application of the news writing robot in American news industry" , Youth Journalist, May, 2016, pp. 99-100.

[3] Gong J, Ren W, Zhang P. "Reflections on application of Machine Writing News", Science & Technology for China's Mass Media, May, 2016, pp. 58-60.

[4] Yu G. "Machine writing news drives new change of the media" , Cover and Edit News, June, 2015, pp. 26-27.

[5] Gong J, Ren W, Zhang P. "An automatic generation method of sports news based on knowledge rules", Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on. IEEE, 2017, pp. 499-502.

[6] Van Dalen A. "The Algorithms Behind the Headlines: how machine-written news redefines the core skills of human journalists", Journalism Practice, vol.6, May, 2012, pp. 648-658

[7] Zong C. "Statistics natural language processing", Tsinghua University press, 2nd, ed. August, 2013.

[8] Zhang J, Chen J. "Summarization of Natural Language Generation",. Application Research of Computers, August, 2006, pp. 1-3.

[9] Aker, A., et al. "Automatic label generation for news comment clusters", In Proceedings of the 9th International Natural Language Generation Conference. 2016. Association for Computational Linguistics.

[10] Nesterenko, L. "Building a system for stock news generation in Russian", In Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016). 2016.

[11] Cullen C, Neill I, Hanna P. "Flexible natural language generation in multiple contexts", Human Language Technology. Challenges of the Information Society. German: Springer Berlin Heidelberg, 2009, pp. 142-153.

[12] Ehud Reiter, SG Sripada, Roma Robertson. "Acquiring Correct Knowledge for Natural Language Generation", Journal of Artificial Intelligence Research, vol. 18, issue. 1, 2011, pp. 491-516.

[13] D Dannélls, N Gruzitis. "Controlled Natural Language Generation from a Multilingual FrameNet-Based Grammar, Springer International Publishing", vol. 8625, 2014, pp. 155-166.

[14] TH Wen, M Gasic, N Mrksic, PH Su, D Vandyke, S Young, "Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems", Conference on Empirical Methods in Natural Language Processing, August, 2015.

[15] Zhou Y, Li Y. "Development principle and application of database" , Tsinghua University press, 2nd, ed, 2013.

[16] Han H. "A review of graph database systems. Computer CD Software and Applications", vol. 23, 2014, pp. 14-15.

[17] Guo X, Zhao S, "Liu J. etc. Knowledge presentation of association rules based on conceptual graphs", Computer Science, vol. 40, August, 2013, pp. 261-265.

[18] Wang Y. "Research on embedded deployment of graphic database Neo4j", Modern Electronics Technique, vol. 22, 2012, pp. 35:36-38.

[19] Gong J, Cao J, Zhang P. "A Knowledge graph-based Content Selection Model for Data-driven Text Generation", International Journal of Reasoning-based Intelligent Systems. vol. 9, March, 2017, pp. 205-209.

[20] Liu Q, Li Y, Duan H, ect. "Knowledge graph construction techniques", Journal of Computer Research and Development, March, 2016, pp. 582-600.

[21] Xu Z, Sheng Y, He L. "Review on Knowledge Graph Techniques Journal of University of Electronic Science and Technology of China", vol. 45, Apirl, 2016, pp. 589-606.

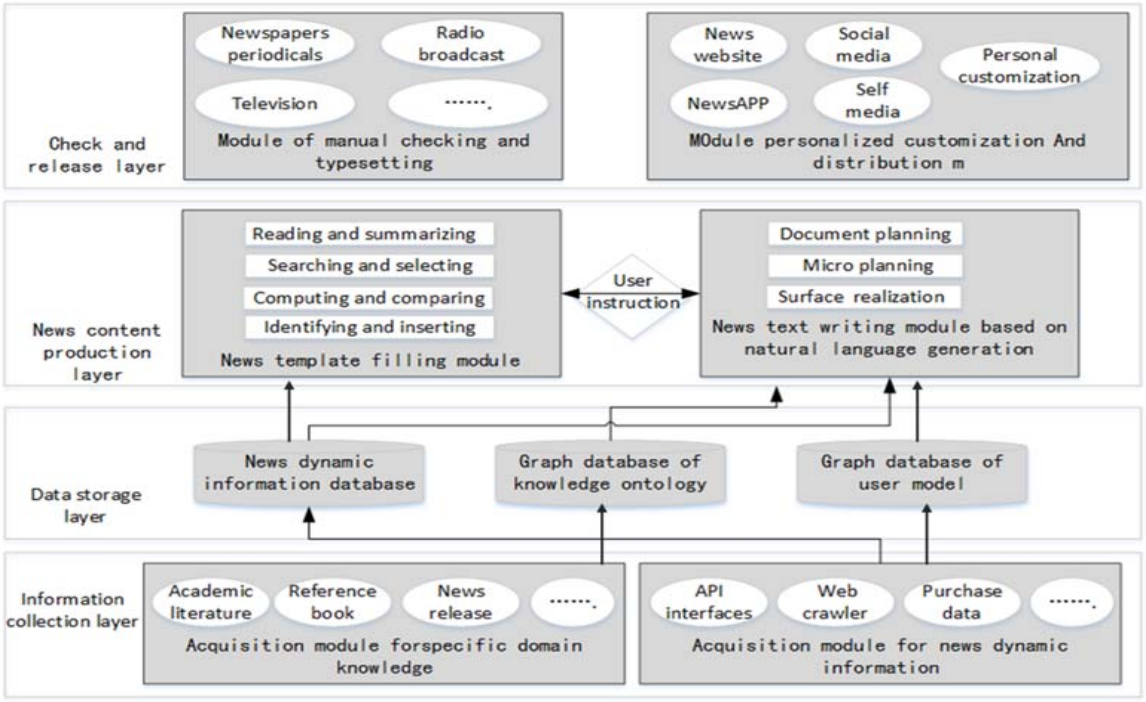

Figure 4. The logical framework of machine news production system