Information retrieval and data mining based on open network knowledge

Deng Ruren[1, a], Wu Yinghuan[2, b]

[1]Jiangxi Science&Technology Normal University, Nanchang, Jiangxi, 330013, P.R. China

[2] Nanchang Institute of Technology, Nanchang, Jiangxi, 330013, P.R. China

[a] 37453477@qq.com, [b]Huanl2016@126.com

Keywords: Open network knowledge; information retrieval; data mining

Abstract. With the progress of science and technology, Internet technology is also developing, changing people's life, but also a huge database. The information in the database contains a complicated and varied information. The construction of the open network information database is an effective means to obtain the information resources, and the information retrieval and data mining is an effective way to resolve the data. This paper discusses the creation of the open network knowledge base, analyzes the information retrieval and data mining based on the open network.

Create an open network knowledge database

The creation of open network knowledge database. Concept extraction, attribute, instance, relationship are the basic elements of the creation of knowledge and information in the open network. In terms of its creation means, it includes two kinds of modes, manual and automatic. First, manual creation. The so-called manual creation refers to the rules that rely on the experts to write, to achieve the collection and integration of the data, and then realize the construction of the knowledge base system. In this mode, emphasize the expert professional, founder of domain knowledge have considerable understanding, in order to ensure that the written rules to meet the requirements, of human and material resources into a larger, HowNet is the typical manually create knowledge base. Second, automatic creation. The so-called automatic creation refers to the artificial intelligence theory, machine learning technology and knowledge engineering technology and the emergence and development of the basis, the use of modern technology to complete the concept extraction and other steps to achieve the creation of the database. Compared with the artificial creation of knowledge base, the automatic creation of the method can save the human and material resources. The automatic creation of knowledge base is also divided into the semi supervised knowledge base and the creation method of supervised knowledge base. In which supervised knowledge base is created by means of the system, the system can be labeled by the web, and then the rules are obtained. Semi supervised database creation method emphasizes the use of machine learning algorithm, through the system defined in advance, to achieve the four elements of the database creation, then, the system will once again to have the elements of the rules to extract, so as to guide the process of the elements of the post, this semi supervised method of building a more widely used, such as Know-ItALL method, TextRunnre method, etc..

Each channel knowledge source integration. Database construction is a systematic project, need to spend a lot of manpower, material resources and financial resources, in order to achieve the sharing and reuse of data resources, it is necessary to achieve the integration of various sources of knowledge. In other words, it is to the database four elements of the existence of duplication, conflict data to clean up and merge, etc.. The integration of the knowledge of the sources of each channel includes two ways of manual integration and automatic integration. Manual integration is relatively time-consuming and laborious, and it is mainly used in the integration of small knowledge base. Automatic integration is based on artificial intelligence, machine learning and other modern technology means, this integration is more convenient and more convenient. At present, YAGO is a typical automatic integration method. The main work of YAGO knowledge base is the integration of Wikipedia knowledge system and WordNet's knowledge system.

The renewal of knowledge information database in open network. With the development and progress of the times, the knowledge system of human beings is constantly changing. Therefore, the update of the database contains two aspects, on the one hand, it is the content of the database, on the other hand, it is a change in the database. For example, the NELL system to the database update is achieved through the way of self correction. As with other systems, the NELL system will continue to extract elements from the Internet, but the correct rate is not one hundred percent, for those who have the wrong relationship and conclusions, NELL system staff will be regularly to find errors in the correction, to achieve the database update.

Information retrieval based on open network knowledge

In the last century 50's, computer technology began to be used in real life, combined with the integration of IT technology and information retrieval technology, and then produced a modern sense of information retrieval". In other words, in the last century before the 50's, the paper media is the main work of information dissemination and storage, information retrieval is also around the paper media. Information retrieval is also known as the Web information retrieval, it has a large number of users, the amount of information, and other characteristics of the professional inconsistent. In the last century 60's, the research on information retrieval technology has made a great breakthrough, and the search engines, such as Google, etc..

Working principle of information retrieval. Under the open network environment knowledge information retrieval system to all kinds of search engines (search engine) based, refers to the so-called search engine is indexing and retrieval system, which can realize the screening of the information in network environment. Information retrieval system basically includes the following three aspects: first, the content of the open network knowledge base is downloaded to the local, and the pretreatment. Second, to achieve the content of the document retrieval. Third, according to the needs of users.

A complete search engine should consist of six parts, in addition to the temporary file database and index database and index database, as well as the index database, as well as the index. Robot is able to achieve the traversal of all the pages of the open network, and through hyperlinks, to achieve automatic roaming pages. On this basis, all the relevant data to be collected and downloaded to the local temporary database. Second, index. The main function of the index is to pre process the documents downloaded from the Robot, and the model is the basis of the model, and the documents are in the form of the document. The index will be stored in the corresponding database to provide convenient for the latter. Third, user interface. The main function of the user interface is to realize the analysis of the user's query information, analysis of the relevant data to meet the requirements of data retrieval. After the retrieval process, the feedback to the user is a retrieval result based on the correlation degree.

Insufficient information retrieval. Existing information retrieval technology, mainly search engine, in the specific use of the results of a more single, and unable to retrieve historical information, but also by a single search engine. At present, most information retrieval process is relatively blind, resulting in the amount of information retrieval is very large, and the user is not enough patience to filter one by one, and it is possible to lose important data information, reduce the retrieval efficiency, but also the waste of data resources. At present, information retrieval search engine, but the amount of information in the open network is also great, a single search engine, it may contain all the information, users may need to search for information retrieval, but it is also possible that all search engines have taken the results are repeated.

Data mining based on open network knowledge

In the light of information retrieval, the focus of data mining is the mining of deep level knowledge. Data mining based on open web site, which mainly includes three aspects: thread mining, relational reasoning and relational forecasting.

Trail mining. The clue is to create the two concepts of the knowledge base and the way to define it. For example, the question of the relationship between the Microsoft Corp and the Apple Corp, as well as other entities related to the acquisition of information, need to be realized through the entity relationship in the information mining. In its actual operation, each point in the knowledge base is defined as a concept, the line between the points represents the relationship between the two, and then through data mining, find out the relevant information. The association method based on the group and the correlation method based on the relation interpretation are the representatives of the method.

Relational reasoning. Relational reasoning refers to the relationship between different entities, such as a, B, C three, B, B, B, B is a C of the mother, then you can determine the potential relationship is a three. At present, the application of relational reasoning is widely used, but also contains many types, mainly based on the rules of FOLL, etc., as well as the Schoenmackers based on the probability graph, etc..

Relationship forecast. Relational forecasting refers to the relationship between entities and concepts, and the main contents of the pre judgment include the occurrence of relationship, the occurrence regularity and the possible changes of relationship. The realization of data mining relationship forecast function is based on machine learning technology, which includes two types of non supervised learning and supervised. For example, in social networks, the possibility of the occurrence of the relationship between the two entities to determine the basis for the realization of the learning method is the basis of the relationship between the problem as a classification problem, in the classification, based on the judgment of the probability of occurrence of the two. Although this approach is widely used, the accuracy of the prediction is still to be improved. The unsupervised learning technology in the forecast, the data and the prior distribution of the exclusion, and thus effectively improve the accuracy of the forecast.

Conclusion

All in all, this paper makes a detailed analysis of the data retrieval and data mining based on the knowledge of the basic development network. In this case, the research work is not deep enough, and it needs further investigation.

References

[1] Y. Tang, X.L. Lu, ye min Luo, etc.. Based on automatic question answering system of information retrieval technology research progress. Computer application, 2012, 28 (11): 2745 - 2748

[2] Diekema, Zhao Jun, Duan Xiangyu,. Question and answer type retrieval technology and evaluation research review. Journal of Chinese information 2015, 19 (3): 1 - 13.

[3] L.W. Chen, Y.S. Feng, D.Y. Zhao. Based on the massive network data relationship extraction. Computer research and development, 2013, 50 (9): 1835 - 1825