An Active Workflow Method for Entity-Oriented Data Collection

Gaoyang Guo(✉)

School of Software, Tsinghua University, Beijing 100084, China
ggy16@mails.tsinghua.edu.cn

Abstract. In the era of big data, people are dealing with data all the time. Data collection is the first step and foundation for many other downstream applications. Meanwhile, we observe that data collection is often entity-oriented, i.e., people usually collect data related to a specific entity. In most cases, people achieve entity-oriented data collection by manual query and filtering based on search engines or news applications. However, these methods are not very efficient and effective. In this paper, we consider designing reasonable process rules and integrating artificial intelligence algorithms to help people efficiently and effectively collect the target data related to the specific entity. Concretely, we propose an active workflow method to achieve this goal. The whole workflow method is composed of four processes: task modeling for data collection, Internet data collection, crowdsourcing data collection and multi-source data aggregation.

Keywords: Data collection · Entity-oriented · Workflow

1 Introduction

In our daily life, people are faced with the need to collect data all the time. For example, football fans may want to get the score data for a current football match. Political lovers may want to obtain the latest political news related to a politician. The fans may want to collect the latest news of their favorite stars. These scenarios can be summed up as a single problem: Given a specific entity such as a football match, a politician, or a star, how can we collect valid data about this entity. We call this problem entity-oriented data collection. In fact, people mainly rely on search engines such as Google or news applications such as RSS to solve their own data acquisition requirements by manual query and filtering nowadays. These methods may be low efficiency. Meanwhile, the data collected by these methods may be scattered. In this paper, we propose an active workflow method to solve this problem.

2 Related Work

Web crawler is a common method for data collection from Internet [8]. A web crawler, sometimes called a spider, is an Internet bot that systematically browses

C. Woo et al. (Eds.): ER 2018 Workshops, LNCS 11158, pp. 76–81, 2018.
https://doi.org/10.1007/978-3-030-01391-2_15

the World Wide Web, typically for the purpose of Web indexing [10]. Web search engines and some other sites use web crawling or software to update their web content or indices of sites' web content [2]. Web crawlers copy pages for processing by a search engine which indexes the downloaded pages so that users can search more efficiently. However, web crawler only returns raw web content, which needs users to further dig out valid data. In this paper, our method can perform a series of subsequent data analysis, which outputs the desired data directly. For example, deep learning models such as LSTM networks can be utilized to extract valuable information from texts [6]. We can also try to integrate data from web based on their community structures [7].

Crowdsourcing is a sourcing model in which individuals or organizations obtain goods and services [1,4]. These services include ideas and finances from a large, relatively open and often rapidly-evolving group of Internet users. It divides work between participants to achieve a cumulative result. In this paper, we also utilize the method of crowdsourcing to collect target data from the real world.

A workflow consists of an orchestrated and repeatable pattern of business activity enabled by the systematic organization of resources into processes that transform materials, provide services, or process information [3,5]. It can be depicted as a sequence of operations, the work of a person or a group, the work of an organization of staff, or one or more simple or complex mechanisms. Petri net [9] is a common method of workflow. In this paper, we borrow the idea of workflow to design our method for entity-oriented data collection.

3 Problem Definition

Entity-Oriented Data Collection. Given a specific entity E, sources S and constraints R, collect and aggregate valid data from S which are related to E and satisfy R.

E is our target entity, which could be a person, a place, an event, a thing or an organization. It specifies what information the user wants. S defines the sources where we collect, which could be some seed URLs or some places in the world. R represents the constraints the target entity E should satisfy, which could be some features expressed in texts, images or videos.

4 Method

In this paper, an active workflow method is proposed to solve this problem. The flow char of this method is shown in Fig. 1. As shown in Fig. 1, the whole workflow method is composed of four processes: task modeling for data collection, Internet data collection, crowdsourcing data collection and multi-source data aggregation. Every process follows some specific process rules. With the help of this workflow method, we can achieve the automation of data collection for specific entities. Given an entity-oriented data collection task defined by the target entity E, the sources S and the constraints R, we first input the task into the process of task modeling for data collection. After that, we obtain an EOSQL statement, which is fed into the process of Internet data collection and the crowdsourcing data collection. Then, the scattered valid data from multi sources are aggregated in the process of multi-source data aggregation. Finally, we collect desired valid data which are related to E and satisfy R. The detailed discussion of each process is as follows.

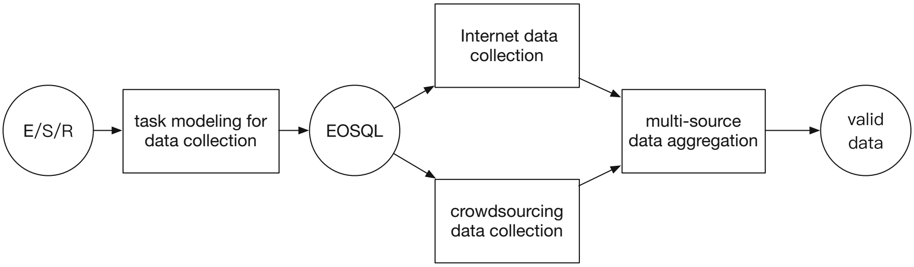
E/S/R
task modeling for data collection
EOSQL
Internet data collection
crowdsourcing data collection
multi-source data aggregation
valid data

Fig. 1. The active workflow method for entity-oriented data collection

4.1 EOSQL

The first process is task modeling for data collection. In this process, an entity-oriented data collection language called EOSQL is designed to model the data collection task. The design of EOSQL is inspired by SQL. The grammar of EOSQL is similar with SQL, which consists of different types of fields and also supports the clause structure. Users can model the task for entity-oriented data collection by simply using an EOSQL statement following the grammar of EOSQL. The grammatical rules of EOSQL are designed as follows:

$$COLLECT\ [E]\ FROM\ [S]\ WHERE\ [R], \tag{1}$$

where E, S, and R are three fields to model the task for entity-oriented data collection. E indicates the target entity the user is interested in. It specifies what information the following processes should collect. S defines the sources where we collect. For example, S could be some seed URLs which are the entry points of the crawler. Also, S could be some places in the world where the target entity exists. R represents the constraints the target entity E should satisfy. For example, R could be some features expressed in texts, images or videos.

EOSQL plays an important role in the whole workflow method. It is the basis of the following processes since it defines the overall goal for the entity-oriented data collection using a uniform language.

4.2 Distributed Data Crawler

The second process is Internet data collection. In this process, we propose an entity-oriented distributed data crawler, which crawls valid data from Internet related to the specific entity. Based on an EOSQL statement, we can obtain E, S and R. The distributed data crawler then crawls the valid data from S which is related to E and satisfies R. In most cases, the crawler deals with the data collection from Internet.

Our crawler mainly consists of two parts: the preliminary crawling and the further extraction. The preliminary crawling is based on Nutch, which is a highly extensible and scalable open source web crawler software project. This part can be divided into 5 stages: distributed URL injection, distributed URL task allocation, distributed URL queue grabbing, distributed URL web page resolution and distributed database update. The further extraction extracts valid information according to E and R. Based on the preliminary crawling, the further extraction utilizes real-time data processing algorithms to recognize and extract the target entity. For example, we can use information extraction methods such as entity recognition and entity relation extraction to collect valid data from texts according to E and R. We can use computer vision techniques such as face recognition and object detection to extract valid data from images or videos according to E and R.

Based on these two parts, our crawler can achieve the tight coupling of crawling and extraction, which can reduce storage overhead and improve collection efficiency.

4.3 Crowdsourcing Collection Model

The third process is crowdsourcing data collection. In this process, an entity-oriented crowdsourcing collection model is proposed to collect raw data from the real world related to the specific entity based on a crowdsourcing algorithm. Based on an EOSQL statement, we can obtain E, S and R. The crowdsourcing collection model then collect the valid data from S which is related to E and satisfies R. In general, this model deals with the data collection from the real world.

The crowdsourcing collection model mainly relies on human resources. It consists of two steps: the shallow crowdsourcing collection and the deep crowdsourcing collection. In the shallow crowdsourcing, the crowdsourcing participants collect large quantity of uncertain data to be confirmed using collection devices. They filter out useless information to reduce the cost of subsequent collection. In the deep crowdsourcing, the model executes the feature analysis, crowdsourcing task division and crowdsourcing task assignment in turn to further process the uncertain data collected from the first step. Among them, crowdsourcing task assignment is the most challenging since this problem is often NP-hard. To solve the problem of task assignment, we should model it as a NP-hard problem first. It is assumed that there are several crowdsourcing participants and several crowdsourcing tasks. Every participant has a cost and a quality for executing each task. The goal is to find a task assignment scheme to minimize the cost and maximize the quality for executing task. Then we need design an approximate algorithm to solve this optimization problem.

Based on these two steps, our crowdsourcing collection model can collect valid data from the real world with a low cost.

4.4 Multi-source Data Aggregation Model

The last process is multi-source data aggregation. In this process, we propose a multi-source data aggregation model to effectively aggregate data collected from the second process and the third process.

After the second process and the third process, we acquire lots of valid data related to the target entity E both from Internet and the real world. However, these data are from different sources and lack effective organization. Moreover, they may have some quality problems such as data redundancy. Therefore, it is necessary to integrate them all using a uniform data format. The main challenge of this process is to design entity matching algorithm for multi-source and multi-domain data. The data collected from the last two processes may have known patterns or my have unknown patterns. To cope with this scenario, we propose a data blocking strategy in hybrid mode, which not only improves the effect of data blocking by using the known patterns, but also have the universality of handling unknown patterns. On the one hand, we put forward a lazy entity matching strategy based on the query perception technology, which conducts entity matching only for the necessary data and only at the necessary time. This strategy can prevent unnecessary computation cost caused by massive Internet data. On the other hand, we also propose an incremental updating strategy for entity matching to deal with the fact that all data sources are constantly updated. Based on these two strategies, the entity matching algorithm is effective and efficient.

Finally, We aggregate all valid data within a uniform data format utilizing the multi-source data aggregation model. Users can acquire desired data which are related to E and satisfy R from the outputs of the last process.

5 Conclusion

In this paper, we introduce a problem called entity-oriented data collection. An active workflow method is proposed to solve this problem. The active workflow method consists of four processes. We design one model and corresponding process rules for each process. In the future, we will continue to improve our workflow method to conduct entity-oriented data collection.

Acknowledgment. This work was supported in part by the National Key Research and Development Program of China (No. 2017YFC0820402), the Intelligent Manufacturing Comprehensive Standardization and New Pattern Application Project of Ministry of Industry and Information Technology (Experimental validation of key technical standards for trusted services in industrial Internet), and the National Natural Science Foundation of China (No. 61373023).

References

1. Buettner, R.: A systematic literature review of crowdsourcing research from a human resource management perspective. In: Hawaii International Conference on System Sciences, pp. 4609–4618 (2015)
2. Corby, O., Dieng-Kuntz, R., Faron-Zucker, C.: Querying the semantic web with corese search engine. In: Eureopean Conference on Artificial Intelligence, ECAI 2004, Including Prestigious Applicants of Intelligent Systems, PAIS 2004, Valencia, Spain, August, pp. 705–709 (2017)
3. Curcin, V., Ghanem, M., Guo, Y.: The design and implementation of a workflow analysis tool. Philos. Trans. Math. Phys. Eng. Sci. **368**(1926), 4193 (2010)
4. Doan, A.H., Ramakrishnan, R., Halevy, A.Y.: Crowdsourcing systems on the world-wide web. Commun. ACM **54**(4), 86–96 (2011)
5. Georgakopoulos, D., Hornick, M., Sheth, A.: An overview of workflow management: from process modeling to workflow automation infrastructure. Distrib. Parallel Databases **3**(2), 119–153 (1995)
6. Guo, G., Wang, C., Chen, J., Ge, P., Chen, W.: Who is answering whom? Finding "reply-to" relations in group chats with deep bidirectional lstm networks. Clust. Comput. **10**, 1–12 (2018)
7. Guo, G., Wang, C., Ying, X.: Which algorithm performs best: algorithm selection for community detection. In: Companion of the The Web Conference, pp. 27–28 (2018)
8. Kobayashi, M., Takeda, K.: Information retrieval on the web. Annu. Rev. Inf. Sci. Technol. **39**(1), 33–80 (2005)
9. Murata, T.: Petri nets: properties, analysis and applications. Proc. IEEE **77**(4), 541–580 (1977)
10. Shaila, S.G., Vadivel, A.: Architecture specification of rule-based deep web crawler with indexer. Int. J. Knowl. Web Intell. **4**(4), 166–186 (2013)