Journal of Digital Convergence
Vol. 19. No. 9, pp. 209-215, 2021

ISSN 2713-6434 / eISSN 2713-6442
https://doi.org/10.14400/JDC.2021.19.9.209

Database metadata standardization processing model using web dictionary crawling

Hana Jeong[1], Koo-Rack Park[2*], Young-suk Chung[3]
[1]Doctoral Student, Department of Computer Engineering, Kongju National University
[2]Professor, Department of Computer Engineering, Kongju National University
[3]Doctor, Department of Computer Engineering, Kongju National University

웹 사전 크롤링을 이용한 데이터베이스 메타데이터 표준화 처리 모델

정하나[1], 박구락[2*], 정영석[3]
[1]공주대학교 컴퓨터공학과 박사과정, [2]공주대학교 컴퓨터공학부 교수, [3]공주대학교 컴퓨터공학과

Abstract Data quality management is an important issue these days. Improve data quality by providing consistent metadata. This study presents algorithms that facilitate standard word dictionary management for consistent metadata management. Algorithms are presented to automate synonyms management of database metadata through web dictionary crawling. It also improves the accuracy of the data by resolving homonym distinction issues that may arise during the web dictionary crawling process. The algorithm proposed in this study increases the reliability of metadata data quality compared to the existing passive management. It can also reduce the time spent on registering and managing synonym data. Further research on the new data standardization partial automation model will need to be continued, with a detailed understanding of some of the automatable tasks in future data standardization activities.

Key Words : Data standardization, Data quality management, Web crawler, Database, Metadata.

요 약 데이터 품질 관리는 최근 중요한 이슈로 자리잡았다. 데이터베이스의 메타데이터 표준화는 데이터 품질관리 방안 중 하나이다. 본 연구에서는 일관된 메타데이터 관리를 위하여 표준단어사전 관리를 지원하는 알고리즘을 제시한다. 해당 알고리즘은 웹 사전 크롤링을 통해 데이터베이스 메타데이터의 동의어 관리 자동화를 지원한다. 또한 웹 사전 크롤링 과정에서 생길 수 있는 동음이의어 판별 이슈를 해결하여 데이터의 정확도를 향상시킨다. 본 연구에서 제안하는 알고리즘은 기존의 수동적 관리에 비해 메타데이터 데이터 품질의 신뢰도를 높인다. 또한 이음동의어 데이터 등록 및 관리에 소비되는 시간을 단축시킬 수 있다. 새로운 데이터 표준화 부분 자동화 모델에 대한 추가 연구는 향후 데이터 표준화 프로세스에서 자동화 가능한 작업을 파악하여 진행되어야 한다.

주제어 : 데이터표준화, 데이터 품질관리, 웹 크롤러, 데이터베이스, 메타데이터

*Corresponding Author : Koo-Rack Park(ecgrpark@kongju.ac.kr)
Received August 31, 2021 Revised September 10, 2021
Accepted September 20, 2021 Published September 28, 2021

1. Introduction

In this era of big data, a lot of data is being generated and collected. In addition, open data policies that disclose data from public institutions are implemented in countries around the world, and data sharing is actively carried out [1-3]. Thus, the amount of data available has increased compared to the past. However, low quality data is prevalent in large databases as data quality control cannot keep up with the rate at which data is generated and collected [4]. because of this, many companies are consuming loss and rework costs due to poor data quality [5-8]. For example, in a 1990 Senate report, the U.S. General Accounting Agency reported that a single agency lost more than $2 billion in federal loans due to poor data quality [9]. As a result, the importance of data quality management has increased. Data standardization, one of the methods of data quality management, exists in many areas. This study deals with the standardization of metadata in databases. Improve data quality by managing the terms used in metadata on a word-by-word basis to consistently manage and maintain metadata. And it can provide reliability and ease of use for the data.

2. Related Works

2.1 Database metadata standardization

The standardization of metadata in a database is to maintain and manage the metadata in consistent terms. In this paper, the table name and column name of the database among metadata are subject to standardization. To manage in a consistent term, a standard term dictionary, a set of terms, is managed [10]. It also separates the terms into words and manages them from the word units. The set of words that are consistent and managed without duplication is called a standard word dictionary. This standard word dictionary does not accept synonyms. This is because allowing synonyms lowers the quality of data standards because duplicate data exists. However, since synonym management in these standard word dictionary is carried out manually, the disadvantage is that the data in the standard word dictionary is unreliable and time consuming.

2.2 Database metadata standardization

A web crawler is a program that automatically discovers and indexes numerous web pages. It also saves links to web pages it has explored for future use. it called spider or web robot, worms [11,12]. Most of these Web crawlers have a link-based crawling strategy [13,14]. Web crawlers can reduce the time it takes to collect data on the web, where vast amounts of data are gathered. Web crawling varies in ways depending on the format of the site. The URL(Uniform Resource Locator) is largely divided into cases where it can be inferred and not [15].

The first is when you can guess the value of a parameter, so if you change only the parameter, you can access the page with the desired information. The second is when the value of the parameter cannot be guessed, or only some of the information to be crawled is provided and the full text is on a different page. Web dictionaries should search for a specific word and select it from the search results list to see the details of the word. Therefore, it has both forms. For search result pages, you can access the page by inserting the words you want to search into a specific URL as parameter values. The search result detail page corresponds to crawl when URL cannot be inferred. Therefore, if you want to crawl a detailed search result page, you should learn the unique number of words in the search result list to obtain the detailed search result page URL.

3. Proposed Method

3.1 Proposed model diagram

It is important to manage the uniqueness of standard words in the standardization of metadata in databases. This is because it determines the quality of metadata standards. Synonyms management of standard words is required to ensure the uniqueness of standard words. However, it takes a lot of time for users to handle it manually and lacks credibility with synonym data. Therefore, a database metadata standardization processing model using Web dictionaries is needed. The proposed model for standardizing database metadata using Web dictionaries is as follows Figure 1.

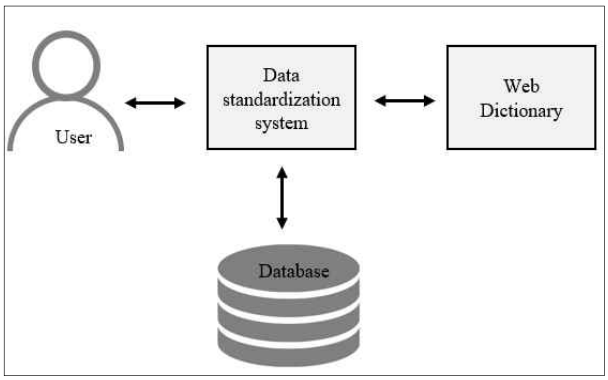

Fig. 1. Diagram of Database Metadata Standardization Processing Model Using Web Dictionary

When a user requests the registration of a standard word in the data standardization system, the synonym is crawled from the web dictionary. This crawling synonym is stored in the database. This collection of synonym data is referred to as synonym dictionary. Standard words are also stored in the database.

The database metadata standard word guidelines are defined as follows.

- Standard words should be defined as nouns.
- Standard words have Korean word, English word, and English word abbreviations as required.
- Standard words should not contain special characters.
- Synonyms cannot exist in duplicate.

3.2 Process for database metadata synonyms using Web dictionaries

Figure 2 is the process of processing synonyms when standardizing metadata using Web dictionaries.

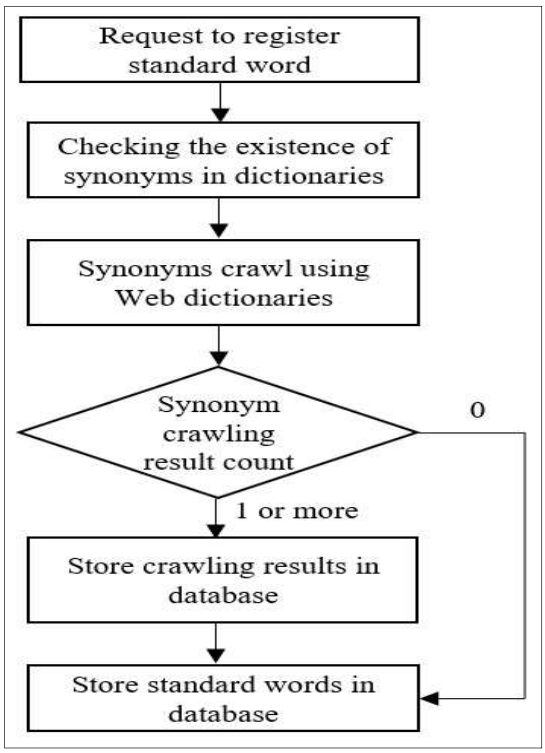

Fig. 2. Process for database metadata synonyms using Web dictionaries

The processing order for each step is as follows. First, the user requests the registration of the standard word by entering the Korean word and English word data of the standard word that they want to register in the dictionary. Second, check if the standard word requested for registration is a word that exists in the synonym dictionary. A synonym dictionary has the value of the synonym Korean words, Chinese word, and URL. Compare the Korean word of the standard word requested for registration with the Korean word of the synonym dictionary to find out if

there is a match. If there is data that matches the Korean word, the user will be checked for matches by providing the Chinese word and URL of the matching word. If the user responds that they agree, the user shall restrict the registration of the standard word by judging that it is synonymous with the standard word dictionary. If not found in the synonym dictionary, it is determined that there is no synonym of the word in the current standard word. Therefore, it allows the registration of standard words. Third, use the Korean dictionary and the English-Korean dictionary to crawl synonyms for newly registered standard words. Fourth, check the number of synonym crawling data. If the synonym has been successfully crawled, there is more than one. If for some reason you haven't crawled the synonym, it's zero. Fifth, if more than one crawling data exists, store it in a synonym dictionary. If the number of synonyms crawled is zero, this step is skipped. Finally, the standard word requested by the user is registered in the standard word dictionary.

3.3 Synonyms crawl using Web dictionaries

The following Figure 3 is an algorithm pseudo-code that differentiates homonyms in Korean from web dictionaries and crawls synonyms.

word_en is the English word for the standard word requested for registration, and word_ko is the Korean word for the standard word requested for registration. In addition, ko_list, en_list, is a two-dimensional arrangement that contains data that crawls the search results of Korean dictionary and English-Korean dictionary. ko_list[n][0] is the Chinese word data of the results of the Korean dictionary search, and ko_list[n][1] is the detailed page URL data of the results of the Korean dictionary search. en_list[n][0] contains a Chinese word for the results of an English dictionary search. en_list[n][1] is English word data in the results of an English dictionary search. Only search results

```
FUNCTION searchCorrectWord{
IF length(ko_list) equal 0
    STOP
END IF

FOR i ← to length[en_list]
    IF en_list[i][1] exist word_en THEN
match_china_word ← en_list[k][0]
        IF match_china_word IS NULL
Match_url ← user_choose_word(ko_list)
                IF match_url IS NULL STOP END IF
                Synonym_list[] ← crollingSynonym(Match_url)
        END IF
    ELSE
Match url ← user choose word(ko list)
        IF match_url IS NULL STOP END IF
        Synonym_list[] ← crollingSynonym(Match_url)
END IF


FOR j ← 0 to length[ko_list]
    IF ko_list[i][0] equal match_china_word THEN
        match_url ← ko_list[i][1]
        synonym_list[] ← crollingSynonym(match_url)
    ELSE STOP END IF
RETURN synonym_list[]
}
```

Fig. 3. Synonym crawl pseudo-code using web dictionary

that fully match word_ko among the pre-search results are taken as values. match_china_word finds the same English word as word_en in the en_list and takes the Chinese word that matches the English word as its English word. match_url finds a word that matches word_ko in the ko_list and takes the URL that matches that word as a value. where URL is the detailed page URL of the word searched in the Internet dictionary. synonym_list is the value obtained by accessing the URL obtained earlier and crawling the synonym of word_ko. The synonym crawling algorithm first checks the length of the ko_list. If the length of the ko_list is zero, the algorithm is interrupted because synonym crawl is impossible. If ko_list is more than 1 length, locate an array row with the same value as word_en as the en_list by a repeating statement. If a row with the same value as word_en exists, take the Chinese word of the row and call it match_china_word. If the match_china_word value does not exist because word_ko does not

have a Chinese word value, give the user a list of ko_list and let them choose the Korean word they want. If a user designates a word in ko_list, he or she connects to the URL, or match_url, that matches the word, and crawls the synonym of word_ko. In addition, if the en_list does not find the same value as word_en, the synonym is crawled by having the user choose as above. If you have successfully obtained the match_china_word value earlier, locate an array row with the same value as match_china_word in the ko_list. If a row with the same value as match_china_word exists, connect to the value of the same row, that is, match_url, and crawl the synonym of word_ko. In other words, the English word entered by the user is searched in the English-Korean dictionary to find the Chinese word of the Korean word matched with the English word, and the Korean word entered by the user is searched in the Korean dictionary. It is a method of crawling synonyms by selecting Korean words that match the Chinese word found earlier in the search results. The reason for using the English-Korean dictionary is that there are many homonyms in Korean. Korean words alone cannot be distinguished When the word entered by the user is a word with homonym.

3.4 Database structure

The following Table 1 is the structure of the standard word dictionary database.

Table 1. Standard word dictionary table

No.	Field	Description	Data Type	NOT NULL	UNIQUE
1	WORD_ID (PK)	Standard word ID	Varchar (15)	Y	Y
2	WORD_KO	Standard word korean name	Varchar (100)	Y	Y
3	WORD_EN_FULL	Standard word english full name	Varchar (500)	Y	Y
4	WORD_NM	Standard word english abbreviation	Varchar (30)	Y	Y

Table 1 WORD_KO, WORD_EN_FULL, and WORD_NM are data entered by the user. WORD_ID is the standard word unique ID value as PRIMARY KEY (PK). The WORD_KO, WORD_EN_FULL, and WORD_NM values are required because they are UNIQUE values and cannot be duplicated and are NOT NULL data.

The following Table 2 is the structure of the synonym dictionary database. SYNONYM_KO, SYNONYM_CHN, and SYNONYM_URL of Table 2 are values obtained by crawling and contain synonym_list values of Figure 3. SYNONYM_ID is synonymous unique ID value as PK. WORD_ID is a forwarder key (FK), referring to the WORD_ID in the standard dictionary table. It is a column to distinguish which standard word is synonymous with.

Table 2. Synonym dictionary table

No.	Field	Description	Data Type	NOT NULL	UNIQUE
1	SYNONYM_ID (PK)	synonym ID	Varchar(15)	Y	Y
2	WORD_ID(FK)	Standard word ID	Varchar(15)	Y	
3	SYNONYM_KO	Synonym korean name	Varchar(500)	Y	
4	SYNONYM_CHN	Synonym chinese name	Varchar(500)	Y	
5	SYNONYM_URL	Synonym detail page URL	Varchar(500)	Y	

4. Results and Discussion

Fifty virtual data were applied to the model proposed in this study. As a result, 84% of the standard words were registered as follows Table 3.

Table 3. Standard word registration rate

Sortation	Number	Rate
Standard word registration	42	84%
Standard word non registration	8	16%

The following Table 4 shows the detailed items according to the status of the standard word registration.

Table 4. Detailed list according to standard word registration status

Sortation	Detailed section	Number	Rate
Standard word registration	Synonym dictionary registration	31	62%
	Synonym dictionary non registration	11	22%
Standard word non registration	Synonym exists	5	10%
	Standard word al reduplication	3	6%

The most basic process of this study is that the standard word is registered, and the synonym dictionary of the standard word is registered. In addition, there are cases in which a word that already exists in the Synonym Dictionary, that is, a standard word that is an synonym, is already registered, and there are cases where a request for duplicate registration of the already registered standard word. It is a normal process not to register standard words at this time. However, the failure of the synonym dictionary to be registered despite the registration of the standard word deviates from the model of this study. In this case, this is the "Standard word registration >Synonym dictionary non registration" section of Table 4. This is 22% of the total data. This data is not suitable for the proposed model, so standard words are registered but synonym data cannot be imported. The reason is that although the English dictionary was searched, the English words of the English dictionary did not match the english word of the standard word did not match.

Another reason is that standard words are compound word or poorly used jargon. It was not a suitable word to search in Korean dictionary or English dictionary. In conclusion, 78% of the data were successfully applied to the proposed process of this study. Consumption time has also been reduced. It took about 35 minutes to manually identify synonyms, and three minutes to proceed using the proposed model. That is, time spent improved by 92%.

5. Conclusion

As a result of the application of the model proposed in this study, 39 out of 50 cases, or 78%, were normally Identify synonyms and registered standard words. and the time spent on standardization also showed 92% improvement. The application of the model of this study makes the identification of synonyms reliable. This ensures the quality of database metadata data. It also saves time in managing data quality. However, if the standard words to be registered are jargon or compound words, they are not found in the general Web dictionary and therefore are not suitable for the model of this study. In the future, we will find out in detail where automation is possible in data standardization activities. And we are going to conduct research on the new data standardization part automation model.

REFERENCES

[1] Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management, 29(4)*, 258-268.
DOI : 10.1080/10580530.2012.716740

[2] Pitt, M. A., & Tang, Y. (2013). What should be the data sharing policy of cognitive science?. *Topics in Cognitive Science, 5(1)*, 214-221.
DOI : 10.1111/tops.12006

[3] Birney, E., Hudson, T. J., Green, E. D., Gunter, C., Eddy, S., Rogers, J., ... & Yu, J. (2009). Prepublication data sharing. *Nature, 461(7261)*, 168-170.
DOI : 10.1038/461168a

[4] Saha, B., & Srivastava, D. (2014, March). Data quality: The other face of big data. In 2014 *IEEE 30th international conference on data engineering* (pp. 1294-1297). IEEE.
DOI : 10.1109/ICDE.2014.6816764

[5] SEnglish, L. P. (1999). Improving data warehouse and business information quality: methods for reducing costs and increasing profits. John Wiley & Sons, Inc.

[6] Haug, A., Zachariassen, F., & Van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management (JIEM), 4(2)*, 168-193.

DOI : 10.3926/jiem.2011.v4n2.p168-193

[7] Kim, W., & Choi, B. (2003). Towards Quantifying Data Quality Costs. *J. Object Technol., 2(4)*, 69-76.

[8] Eppler, M., & Helfert, M. (2004, November). A classification and analysis of data quality costs. *In International Conference on Information Quality* (pp. 311-325).

[9] Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A framework for analysis of data quality research. *IEEE transactions on knowledge and data engineering, 7(4)*, 623-640.
DOI : 10.1109/69.404034

[10] Lawrence, R., & Barker, K. (2001, March). Integrating relational database schemas using a standardized dictionary. *In Proceedings of the 2001 ACM symposium on Applied computing* (pp. 225-230).
DOI : 10.1145/372202.372327

[11] Shrivastava, V. (2018). A methodical study of web crawler. *Vandana Shrivastava Journal of Engineering Research and Application, 8(11)*, 01-08.
DOI : 10.9790/9622-0811010108

[12] Dhenakaran, S. S., & Sambanthan, K. T. (2011). Web crawler-an overview. I*nternational Journal of Computer Science and Communication, 2(1)*, 265-267.

[13] Pal, A., Tomar, D. S., & Shrivastava, S. C. (2009). Effective focused crawling based on content and link structure analysis. arXiv preprint arXiv:0906.5034.

[14] Jamali, M., Sayyadi, H., Hariri, B. B., & Abolhassani, H. (2006, December). A method for focused crawling using combination of link structure and content similarity. I*n 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)* (pp. 753-756). IEEE.
DOI : 10.1109/WI.2006.19

[15] You, F., Gong, H., Guan, X., Cao, Y., Zhang, C., Lai, S., & Zhao, Y. (2018, August). Design of data mining of WeChat public platform based on Python. *In Journal of Physics: Conference Series, 1069(1),* p. 012017. IOP Publishing.
DOI : 10.1088/1742-6596/1069/1/012017

정 하 나(Ha-na Jeong) [정회원]

· 2019년 2월 : 공주대학교 소프트웨어공학과(공학사)
· 2021년 2월 : 공주대학교 대학원 멀티미디어공학과(공학석사)
· 2021년 3월 ~ 현재 : 공주대학교 대학원 컴퓨터공학과 박사 과정 재학중
· 관심분야 : 데이터베이스, 데이터품질관리, IT 컨버전스
· E-Mail : konghanaj@gmail.com

박 구 락(Koo-Rack Park) [정회원]

· 1986년 2월 : 중앙대학교 전기공학과(공학사)
· 1988년 2월 : 숭실대학교 전자계산학과(공학석사)
· 2000년 2월 : 경기대학교 전자계산학과(이학박사)
· 1991년 ~ 현재 : 공주대학교 컴퓨터공학부 교수
· 관심분야 : IT 컨버전스, 정보통신, 머신러닝, 전자상거래
· E-Mail : ecgrpark@kongju.ac.kr

정 영 석(Young-Suk Chung) [정회원]

· 2009년 2월 : 공주대학교 멀티미디어공학과(공학석사)
· 2013년 2월 : 공주대학교 컴퓨터공학과(공학박사)
· 2009년 ~ 현재 : 공주대학교 외래강사
· 관심분야 : 데이터베이스, 시뮬레이션, 인공지능
· E-Mail : merope@kongju.ac.kr