Research Paper

A Bootstrapping-based Method to Automatically Identify Data-usage Statements in Publications

Qiuzi Zhang, Qikai Cheng, Yong Huang & Wei Lu†

School of Information Management, Wuhan University, Wuhan 430072, China

Citation: Qiuzi Zhang, Qikai Cheng, Yong Huang & Wei Lu (2016). A Bootstrapping-based Method to Automatically Identify Data-usage Statements in Publications.
Received: Jan. 21, 2016
Revised: Feb. 19, 2016
Accepted: Feb. 26, 2016

Abstract

Purpose: Our study proposes a bootstrapping-based method to automatically extract data-usage statements from academic texts.

Design/methodology/approach: The method for data-usage statements extraction starts with seed entities and iteratively learns patterns and data-usage statements from unlabeled text. In each iteration, new patterns are constructed and added to the pattern list based on their calculated score. Three seed-selection strategies are also proposed in this paper.

Findings: The performance of the method is verified by means of experiments on real data collected from computer science journals. The results show that the method can achieve satisfactory performance regarding precision of extraction and extensibility of obtained patterns.

Research limitations: While the triple representation of sentences is effective and efficient for extracting data-usage statements, it is unable to handle complex sentences. Additional features that can address complex sentences should thus be explored in the future.

Practical implications: Data-usage statements extraction is beneficial for data-repository construction and facilitates research on data-usage tracking, dataset-based scholar search, and dataset evaluation.

Originality/value: To the best of our knowledge, this paper is among the first to address the important task of automatically extracting data-usage statements from real data.

Keywords Data-usage statements extraction; Information extraction; Bootstrapping; Unsupervised learning; Academic text-mining

† Corresponding author: Wei Lu (E-mail: weilu@whu.edu.cn).

JDIS
Journal of Data and Information Science
Vol. 1 No. 1, 2016
pp 69–85
DOI: 10.20309/jdis.201606
http://www.jdis.org

1 Introduction

Scientific data function as evidence of veracity in scientific research argumentation (Parsons, Duerr, & Minster, 2010). As the sharing and reuse of scientific data is integral to research progress, effective ways in which data can be accessed, explored, compared, organized, and exchanged among and across academic fields and sub-fields are key to the acceleration of problem-solving and disciplinary advancement (Aalbersberg, Dunham, & Koers, 2013; Chao, 2011; Mooney & Newton, 2012; Piwowar & Chapman, 2008). The Text REtrieval Conference (TREC) and competition established in 1992, cosponsored by the US National Institute of Standards and Technology and US Department of Defense, supports research within the information retrieval community by providing the infrastructure for large-scale evaluation of text retrieval methodologies for use in industry and academia. TREC requires participants to adopt a unified dataset released on the same track to collaboratively promote the resolution of issues related to information retrieval. The ArrayExpress public database archive[①], located at the European Bioinformatics Institute, has also become a major resource for the research community to reuse data from microarray and high-throughput functional genomics experiments.

As the quantity of digital scientific data becomes extraordinary, identifying, mining, and organizing valuable information from these data sources becomes a challenge. Ever-increasing data repositories, especially in quickly changing and increasingly important fields such as biology, medicine, and earth sciences, are being developed and deployed (Robinson, Jiménez, & Torres, 2015; Torres, Martín, & Fuente, 2014). Data-related research, such as data-usage tracking (Konkiel, 2013; Mayernik, 2013), motivations and influences of data-sharing (Piwowar, 2011; Piwowar & Chapman, 2008; Piwowar & Vision, 2013), and dataset evaluation is flourishing.

These and related studies are committed to exploring the value of scientific data as a kind of emerging academic resource. The basis of the above endeavors and much research, however, has been to identify effective ways to extract data-usage statements (DUS), which specify how scientific data are obtained, processed, and utilized by author(s), using tools that are largely semi-automatic or human-intensive. They include: (1) retrieving literature from academic databases through manually formulated queries and then filtering them by manual inspection (Belter, 2014; Piwowar, Carlson, & Vision, 2011); (2) constructing rules for dataset identification using metadata recorded in data repositories, such as DataOne[②], GEO[③], and ArrayExpress (Mayernik, 2013; Parsons, Duerr, & Minster, 2010; Piwowar &

① http://www.ebi.ac.uk/arrayexpress/
② https://www.dataone.org/
③ http://www.ncbi.nlm.nih.gov/geo/

Chapman, 2008); and (3) machine learning (Névéol, Wilbur, & Lu, 2011; Piwowar & Chapman, 2008). Although simple to implement, these approaches are human-intensive and can only be applied to limited data.

In this paper, we propose a bootstrapping-based unsupervised method to automatically extract DUS without manual intervention, which is independent on research field, in other words, can be easily adapted to different academic areas. Satisfactory results are obtained when the method is applied to computer science, a typical data-driven research field, such as using data to test hypotheses or verify algorithms. The rest of the paper is organized as follows. In Section 2, we present related work. Section 3 elaborates the fundamentals and procedures of our approach. Section 4 illustrates our experimental process and reports the evaluation results. Section 5 presents concluding remarks.

2 Literature Review

Data-usage in academic literature can be divided into two major categories: (1) using data created or introduced by others, i.e. data-reuse, and (2) using first-hand data created by observation, measurement, recording, searching, etc. In this paper, we extract DUS using both of these data-usage patterns.

Digital Object Identifier (DOI), database accession number, names of data repositories, and other relevant references constitute important indications for extraction of data-related statements. These features are often used to formulate queries or construct rules for pattern recognition. Piwowar, Carlson, and Vision (2011) collected research papers in academic search engines using DOI as search queries, and manually examined whether some specific data or datasets were used in the retrieved literature. Belter (2014) studied the usage of three well-known oceanographic data collections. In his study, citations of these data collections are estimated by querying databases using their names as queries. Some attempts were carried on to retrieve articles whose authors share their data by determining whether there is a link to a certain data repository (Piwowar, 2011; Piwowar & Vision, 2013).

Machine-learning methods have also been used to allow the possibility of extracting data-related statements. Piwowar and Chapman (2008) employed both machine learning and pattern matching to determine whether or not authors in the biomedical field disclosed their datasets in their papers. Névéol, Wilbur, and Lu (2011) constructed a support vector machine (SVM) classifier to implement automatic recognition of data deposition statements in medical literature, but used training and test sets from manual labeling. Given the lack of existing data repositories and high cost of manual annotation, unsupervised methods of data extraction offer major advantages. For example, Boland et al. (2012) used a bootstrapping method to identify references to datasets in research papers. Although their method achieved satisfactory performance, some problems remained. In Boland's method, judgment of the validity of the pattern relied on a manual set threshold, where the number of initial seed words was designated as 1.

3 Methodologies

3.1 Problem Description

Data-usage statements (DUS) refer to those statements describing the name, source, structure, compositions, or application of the dataset used in academic literature. The smallest unit of a statement is a single sentence divided by commas. Some positive and negative examples are given in Table 1. DUS identification is achieved by extracting these statements from academic publications. Specifically, given a research paper P, which consists of a set of sentences $S = \{S_1, S_2, \ldots, S_n\}$, identification of DUS is the process of finding the actual subset S_{data} from S, and $\{\forall s \in S_{\text{data}}$, s is a data-usage statement$\}$. In this paper, the DUS extraction problem is solved as a classification problem, i.e. for a sentence S_i, the task is to determine whether or not it is a data-usage statement. As shown in Table 1, the components of a data-usage statement are stable. The core elements include: (1) A word/phrase referring to data objects, which function as a kind of indication for extraction (hereinafter referred to as "data_clue"), and (2) a sequence of words reflecting the relationship between the data object and the corresponding article (noted as "data_pattern"). In this manner, DUS can be abstracted to a pair of a data_clue word and a data_pattern. Extracting such statements can thus be solved by extracting a < data_clue, data_pattern > pair, which can be obtained through a bootstrapping process.

Table 1. DUS examples.

Statements	Positive (☑) OR negative (☒)
In our experiments, the experimental subset contains 1,552 images selected from the GT database and the FERET databases.	☑ The name, source, and compositions of data
The large-scale database contains 93,638 images captured from 9,668 palms of 4,834 individuals, in which 4–10 images are collected for each palm.	☑ The source and compositions of data
Consequently, both of the two experimental subsets contain 1,200 samples for training and 1,200 samples for testing.	☑ Data compositions and application
In order to show the robustness over short noisy intervals and satisfy the two defined semantics R1 and R2, we generate two completely separated clusters, C1 and C2, using two disjoint interval sequences, Q1 and Q2, and add the synthetically generated short noisy intervals marked in red.	☒ Algorithm description
Each group contains 10 subjects.	☒ Experiment participants
The average training time of the repeated random sub-sampling validation is 1.83 × 30 = 54.9 s, and that of the CBE cross-validation is 1.84 × 5 = 9.2 s.	☒ Experiment process

3.2 Theoretical Framework

3.2.1 Bootstrapping Framework

As words of the same type are contextually similar to each other, if we set a few words as the starting point for searching, accompanied by their common contextual features, we can identify more words that possess similar contextual features. By repeating these processes, we can find more and more words that are similar to each other. What is needed to achieve this goal is the selection of probable representative data_clue words as the initial seed words and data_patterns of these data_clues. In this case, an unsupervised bootstrapping method is proposed (Figure 1), which is applied to acquire a set of data_clue and data_pattern pairs sharing similar features, i.e. a list of < data_clue, data_pattern >.

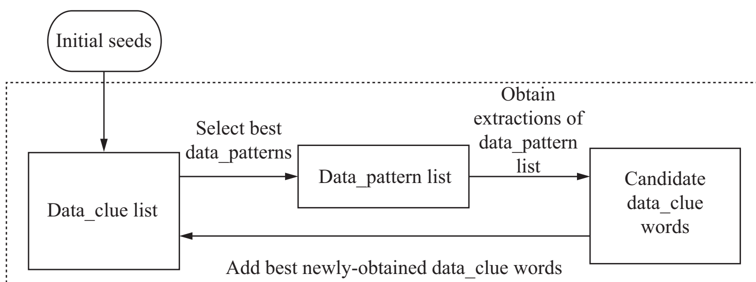

Figure 1. Bootstrapping framework for extraction.

3.2.2 Seed-word Selection

Seed-word selection is the only process that requires manual intervention in our method. The quality of seed words will directly affect the performance of the extraction method. In this paper, three seed-word selection strategies are chosen for implementation:

1) Selecting both the names of a few well-known datasets and a category of general indicative words, such as “dataset,” as seed words, referred to as COM-SEED;
2) Selecting the names of a few well-known datasets as seed words, referred to as SPE-SEED;
3) Selecting a category of general indicative words, such as “dataset,” as seed words, referred to as GEN-SEED.

In order to compare the performances of different strategies, we respectively conduct extraction experiments with each of the three different strategies for seed-word selection.

3.2.3 Pattern Construction

A pattern is used to describe the structural features of the target sentence to be extracted. Normally, the stronger the generalizability of a pattern, the wider the scope of the sentences covered by it; conversely, the stronger the representativeness of a pattern, the narrower the scope of the sentences covered by it. Synthetically, considering both the generalizability and representativeness of patterns, we believe that components of a pattern should at least include the core part of a complete sentence, i.e. the predicate part. Meanwhile, given that the seed words, which are almost exclusively nouns or noun phrases, usually occur in the subject or object part, we choose to construct two types of patterns:

1) Subject part + predicate part, particularly dealing with circumstances in which seed words occur in the object part;
2) Predicate part + object part, particularly dealing with circumstances in which seed words occur in the subject part.

Table 2 provides some examples of the two types of patterns and presents their extraction performance.

Table 2. Exemplifications of pattern construction.

Pattern	Sentences covered by this pattern and the extracted data_clue words
Consists of # samples	**The breast cancer set** consists of 569 samples with 357 benign and 212 malignant. Dataset 1 is referred to as **Char250**, which has 250 samples per category for lower and upper cases, respectively; dataset 2 is referred to as **Char1000**, which has 1,000 samples per category for lower and upper cases, respectively. (Please note this pattern occurs twice here.)
We perform experiments	To assess the ability of the proposed clustering algorithm for classifying the shape classes, we perform experiments on **an increasing number of shapes** in the two Aslan and Tari datasets. We perform our experiments on **a real-estate system** with real-life house dataset used in.

3.2.4 Identification of DUS based on Data_clue and Data_pattern

In computer science, a single data-usage relationship pattern can be applied to varying data objects, e.g. different face-recognition datasets may all serve as a training dataset for machine learning in different articles. Moreover, a dataset entity can have multiple applications, e.g. the ClueWeb09 dataset has regularly been used for search-result ranking, query sub-topic mining, relevance evaluation of retrieval systems, and many other types of experiments. We therefore presume that any combination of one data_clue and one data_pattern adopted from their respective result list has the potential to be involved in a DUS. If a sentence contains at least one data_clue and at least one data_pattern belonging to its own result lists, it can be identified as a DUS.

3.3 Procedures

3.3.1 Bootstrapping Process

In each iteration, a number of data_clue words and data_patterns will be, respectively, added to their corresponding final list if their own score has exceeded the current threshold. The score can be interpreted as the relative probability of a data_clue word or a data_pattern being regarded as valid, based on currently available evidence.

As illustrated in Figure 1, the bootstrapping process is triggered by adding the original seed words to the seed pool, and the procedures will be performed as below (define the current iteration as *i*, and the maximum number of iterations as *MAX*):

1) Obtain all of the patterns in the dataset of research papers;
2) Calculate the scores of each pattern, and add the patterns whose scores are within TOP (20+*i*) into the data_pattern list.

The pattern score is calculated with Equation (1):

$$\mathrm{Score}(P) = \frac{F \times \log_2 F}{N}. \tag{1}$$

The above equation was originally used in Riloff's study (1996) for extraction pattern learning. *N* refers to the length of the current data_clue list, and *F* refers to the distribution value in the current data_clue list, i.e. how many unique data_clue words can be extracted through this pattern. In this paper, only patterns which can extract at least one data_clue from the current list are calculated, i.e. on the condition that $F > 1$.

The valid range set to Top (20+*i*) is to avoid situations from occurring wherein no new valid pattern can be generated during the iterative process. In this setting, the range of valid patterns is widened as the number of iterations increases.

3) Make use of the patterns in the current data_pattern list to extract candidate data_clue words;
4) Calculate scores of each candidate data_clue word, and add candidate words whose scores are within the top five into the data_clue list. The word score is calculated with Equation (2):

$$\mathrm{Score}(W) = \frac{\sum_{j=1}^{P} \log_2 (F_j + 1)}{P}. \tag{2}$$

The above equation was first used by Thelen and Riloff (2002) for semantic lexicon learning. *P* refers to how many unique patterns can be used to extract this candidate word, and F_j refers to how many of the total words can be extracted by this pattern that have already been validated, i.e. those that are a member of the current data_clue list.

5) If $i < MAX$, return to Step 2; otherwise, stop the iteration.

3.3.2 Identification of DUS

After the bootstrapping process, a collection of < data_clue, data_pattern > pairs can be generated by randomly selecting a data_clue word from the final data_clue list, and meanwhile randomly selecting a data_pattern from the final data_pattern list to make a combination, which is referred to as "pair_set." If any single sentence contains components that are in accordance with any certain pair existing in the pair_set, it is identified as a DUS.

4 Results

4.1 Data Collection and Preprocessing

Full-text articles from 116 computer science journals published between 2000 and 2014 in ScienceDirect were used for evaluation. We manually collected data and transferred the format of articles from HTML to well-developed XML. In order to facilitate pattern acquisition and noise reduction, the following pre-processes were conducted, resulting in 6,586,852 relations in total:

1) Remove equations in the body of the articles;
2) Remove all of the XML elements whose headings do not contain "result/results," "experiment/experiments," or "evaluation";
3) Extract relations from the texts in the form of a triple (subject, predicate, and object) through a program called ReVerb[④] (Fader, Soderland, & Etzioni, 2011) designed to automatically identify and extract binary relationships from English sentences when the target relations cannot be specified in advance and speed is important. An example is as follows:

   Sentence: *Bananas are an excellent source of potassium*.
   Relation output: (*bananas*, *be source of*, *potassium*).

Consequently, the final data collection for extraction is a collection of sentences embedded in experiment-related sections affiliated with triple relations extracted from them, referred to as CSExperiment-triple hereinafter. The whole data collection is split into two parts: sentences derived from articles published during 2000–2013, and those published in 2014. The former part is for the main extraction experiment,

[④] https://github.com/knowitall/reverb

and the latter is for pattern extensibility evaluation. It should be noted that the number of relations embedded in one single sentence may be greater than one.

4.2 Extraction Experiments

According to the three different strategies designed for seed-word selection, we performed a series of extraction experiments on the CSExperiment-triple (2000–2013) data collection, and each time the maximum number of iterations was set to be 300. As the iteration progressed, we regularly inspected the extraction results and found that the performance of the SPE-SEED strategy was far from ideal. For this reason, we decided to abolish the SPE-SEED experiment and only report the results of the other two seed-selection strategies: COM-SEED and GEN-SEED. The final yield of the entire iterations includes a list of data_clue words and a list of data_patterns, both accompanied by the final scores of each member. Table 3 shows the initial seed words used in the practical experiments.

Table 3. Initial seed words.

Seed-selection strategy	COM-SEED	GEN-SEED
Initial seed words	trec # kdd cup trec wall street journal the # kdd cup dataset corpus	data dataset corpus data set

Note. "#" refers to a specific year. COM-SEED refers to the strategy of selecting both the names of a few well-known datasets and a category of general indicative words as seed words. GEN-SEED refers to the strategy of selecting a category of general indicative words as seed words.

With final lists of data_clue words and data_patterns available, we accumulated all sentences which contained at least one data_clue word and were simultaneously in line with at least one data_pattern in CSExperiment-triple (2000–2013) data collection, which correctly generated the target sentence collection conforming to our definition of DUS. The results are displayed in Table 4, in terms of the total number of data_clue words and data_pattern concerning different seed-selection strategies, and the total number of DUS was counted by relation triples.

Table 4. Elementary statistics on extraction results.

Seed-selection strategy	Pattern	Seed number	Pattern number	Statement number
COM-SEED	Predicate + Object	14,000	670	29,722
	Subject + Predicate	5,105	596	11,869
GEN-SEED	Predicate + Object	18,235	404	35,711
	Subject + Predicate	5,530	334	11,247

Note. COM-SEED refers to the strategy of selecting both the names of a few well-known datasets and a category of general indicative words as seed words. GEN-SEED refers to the strategy of selecting a category of general indicative words as seed words.

4.3 Evaluation and Results Analysis

We believe that a thorough evaluation should be considered in two facets: (1) the performance of the proposed method on extracting DUS in the field of computer science, and (2) the extraction extensibility of the data_patterns in the final list.

4.3.1 Extraction Precision

(i) Evaluation measures: Random-selected samples of the experimental results are applied for evaluation in order to achieve a reasonable balance between the feasibility and reasonability of manual evaluation. We separately selected 300 statements extracted under the guideline of pairwise combinations of two seed-selection strategies and two pattern-construction strategies, which resulted in a total of 1,200 statements for manual judgment. The authors invited four master candidates in information science to make the judgments through tagging every statement according to whether or not it is within the realm of extraction. The operation instruction is as follows: “If a statement meets our extraction target, give a Yes tag; if not, give a No tag; and if it is not clear enough to determine, give an Unknown tag.” The majority of Unknown statements are those which only mention a dataset, but do not illuminate how the dataset relates to the corresponding article. To deal with all of the Unknown statements, the authors scrutinize all of the original full-text articles containing Unknown statements to make a final judgment of either Yes or No. The number of Yes statements is denoted by R, the total number of statements T, and the extraction precision is calculated as Equation (3):

$$\text{Precision} = R \,/\, T\,. \tag{3}$$

(ii) Results analysis: As can be seen from Table 5, different seed-selection strategies and pattern categories affect extraction precision to varying extent, in which both precisions under the pattern in the form of “Predicate + Object” are above 90%. Regardless of what pattern category is adopted, the COM-SEED strategy performs significantly better than GEN-SEED, in terms of precision.

Table 5. Precision of statement extraction from CSExperiment-triple (2000–2013).

Seed-selection strategy	Pattern	Precision (%)
COM-SEED	Predicate + Object	96.34
	Subject + Object	69.67
	Overall	83.01
GEN-SEED	Predicate + Object	95.34
	Subject + Predicate	37.00
	Overall	66.17

Note. COM-SEED refers to the strategy of selecting both the names of a few well-known datasets and a category of general indicative words as seed words. GEN-SEED refers to the strategy of selecting a category of general indicative words as seed words.

Specifically, for different seed-selection strategies, the pattern form of "Predicate + Object" performs substantially better than that of "Subject + Predicate," which is consistent with its structural properties as a human language. Assuming that we intend to generate a well-formed sentence, it will be much easier to seek out an eligible object for a given combination of subject and predicate than to seek out an eligible subject for a given combination of predicate and object. In other words, when the data_clue words are embedded in the subjective section and will be extracted through patterns in the form of "Predicate + Object," the connection between them and the initial seed words is closer and more stable. Conversely, when the data_clue words are embedded in the objective section and extracted through patterns in the form of "Subject + Predicate," the newly added data_clue word will be more prone to deviate from the scope of the initial seed words during the process of iteration.

Given the fact that the extraction precision under the COM-SEED strategy is much greater than that under the GEN-SEED strategy, it is logical to deduce that the specific dataset names added in the initial seed words will confine the contexts of candidate words within the target extraction range, which reduces noise caused by general indicative words, such as "data" or "dataset."

4.3.2 Extensibility of Patterns

(i) Evaluation measures: Extensibility of patterns is the effectiveness of extracting DUS from articles which do not belong to the original data collections, using patterns existing in the final pattern list which were previously acquired by the iterative process on the original article collection. The articles to be extracted include those within the same research field that do not exist in the original data collection, and articles which belong to other research fields. Taking the objective of the present paper into account, only extensibility within the same field is evaluated.

To evaluate the within-field extensibility of patterns, we randomly select 25 unique articles from the CSExperiment-triple (2014) data collection to create the evaluation dataset, which contains 2,015 sentences in total. A golden standard of 487 sentences concerning data-usage is generated through manual annotation from the evaluation dataset word-for-word. Any sentence which meets at least one data_pattern in the final list is automatically extracted from the evaluation dataset to form the results collection, which are then compared with the golden standard.

Extensibility evaluation is achieved through comparing the results collection with the golden standard. Counted by sentences divided by periods, the number of all sentences in the results collection is denoted by R_n; the number of common sentences both in the results collection and in the golden standard is denoted by M_n; and the number of sentences of the golden standard is denoted by S_n. We use precision (accuracy of extension), recall (coverage of extension), and F-measure (overall extensibility) for evaluation. These three measures are calculated with Equations (4)–(6):

$$\text{Precision} = \frac{M_{\text{n}}}{R_{\text{n}}}, \tag{4}$$

$$\text{Recall} = \frac{M_{\text{n}}}{S_{\text{n}}}, \tag{5}$$

$$F\text{-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{6}$$

(ii) Results analysis: Figures 2 and 3 show the extensibility of pattern changes over the process of iteration in different ways under different seed-selection strategies.

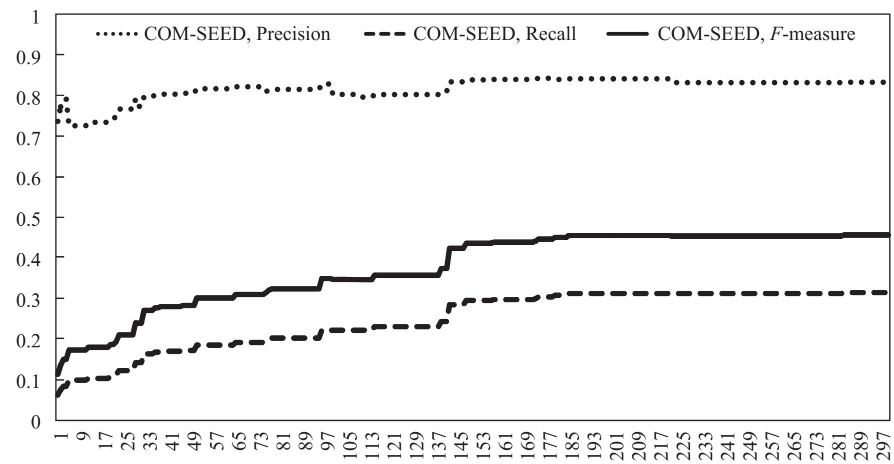

Figure 2. Extensibility of pattern changes over the process of iteration under COM-SEED. COM-SEED refers to the strategy of selecting both the names of a few well-known datasets and a category of general indicative words as seed words.

The recall rate exhibits a noticeable tendency in the following way: within a certain range, as the number of iterations increases, the number of valid patterns increases and the recall rate rises accordingly. As the number of iterations exceeds a certain range, however, the number of valid patterns extracted by the bootstrapping method gradually stabilizes, so does the recall rate. In terms of precision, an evident distinction between COM-SEED and GEN-SEED is recognized and can be described as follows: under the former strategy, the precision is capable of maintaining a high

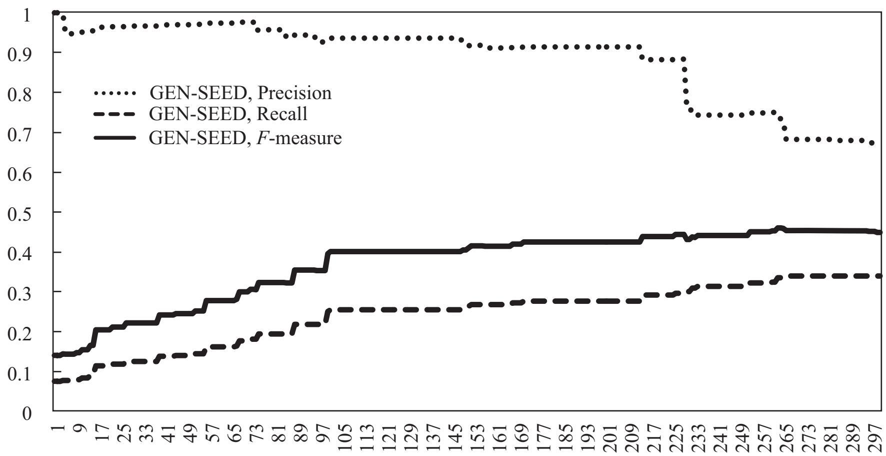

Figure 3. Extensibility of pattern changes over the process of iteration under GEN-SEED. GEN-SEED refers to the strategy of selecting a category of general indicative words as seed words.

level, whereas under the latter strategy, a clear decreasing trend appears after a temporary stability in the early iteration period. The preceding phenomenon is consistent with our conclusion made in the previous section, in which the integration of the two types of words as initial seed words (i.e. COM-SEED) will refine the contexts and make them better accordance with the extraction range.

As shown in Figure 4, for the "Predicate + Object" type of pattern, the difference between the two seed-selection strategies in the performance on recall is insignificant: the precision rate under GEN-SEED exhibits a sudden decrease after a certain number of iterations in which COM-SEED remains stable.

As seen in Figure 5, regarding the "Subject + Predicate" type of pattern, the GEN-SEED strategy possesses advantages over the COM-SEED strategy both in precision and recall, although it is slightly inadequate in the stability of precision. It can thus be observed that the COM-SEED strategy is more suitable for extracting the "Predicate + Object" type of pattern, where the GEN-SEED strategy is more suitable for extracting the "Subject + Predicate" type of pattern. Figure 6 presents the performance of the extensibility of patterns in a special circumstance, in which only the COM-SEED strategy is executed to extract patterns in the form of "Predicate + Object," and meanwhile only the GEN-SEED strategy for "Subject + Predicate." From Figure 6, on the experiments of within-field extensibility of patterns, the performance of our method on precision reaches a sufficiently high level, and an acceptable level of recall can be guaranteed at the same time.

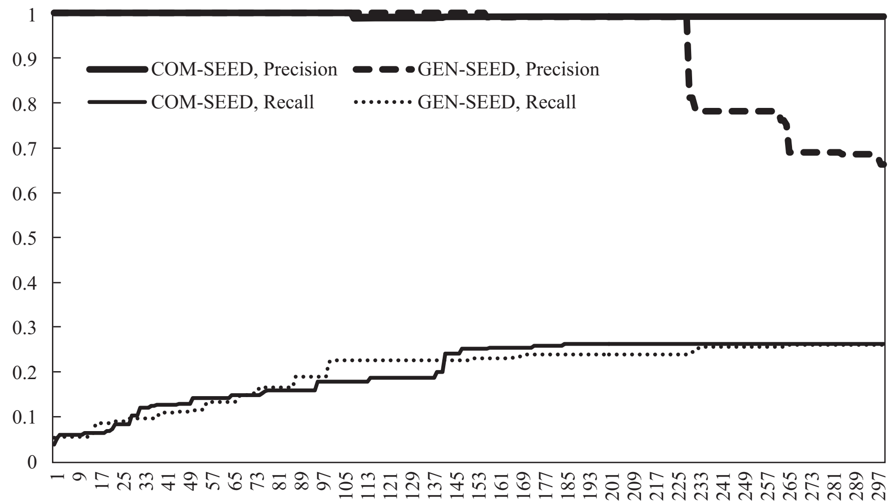

Figure 4. Extensibility of patterns in the form of "Predicate + Object" changes over the process of iteration. COM-SEED refers to the strategy of selecting both the names of a few well-known datasets and a category of general indicative words as seed words. GEN-SEED refers to the strategy of selecting a category of general indicative words as seed words.

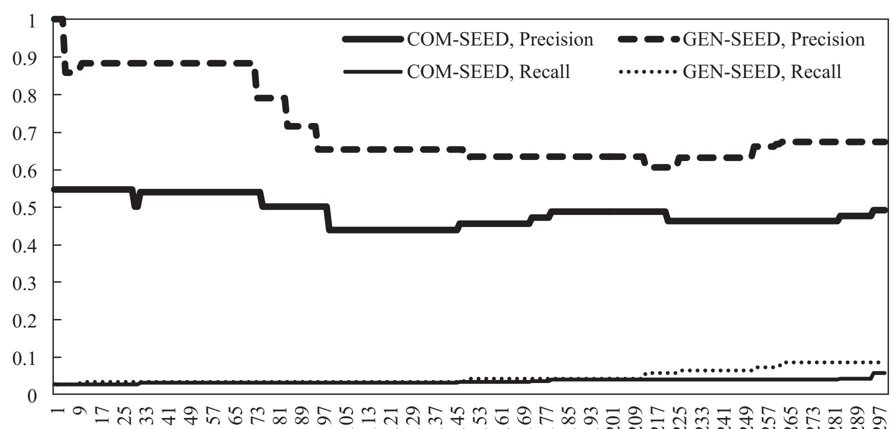

Figure 5. Extensibility of patterns in the form of "Subject + Predicate" changes over the process of iteration. COM-SEED refers to the strategy of selecting both the names of a few well-known datasets and a category of general indicative words as seed words. GEN-SEED refers to the strategy of selecting a category of general indicative words as seed words.

5 Conclusion and Future Work

Scientific data-usage is currently one of the most important academic tools for managing data-driven research that is quickly developing and flourishing. Studies of its motivations and effects constitute critical efforts to explore the behavioral

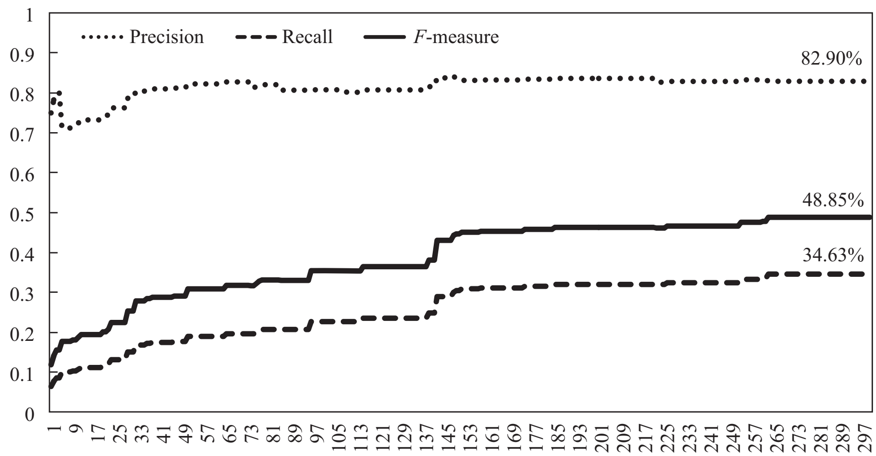

Figure 6. Extensibility of pattern changes over the process of iteration under an optimum combination of seed-selection strategy and pattern-construction strategy.

characteristics of scientific data-usage, making it possible to better utilize the value of data and effectively serve increasing numbers of researchers. Yet due to the complexity of the diverse behaviors and lack of common standards for using data, most studies still follow traditional research concepts that focus on literature citation to perform data citation analysis. Unlike such studies, this article conducts a preliminary investigation of this issue from the perspective of extracting DUS.

The proposed automatic extraction method based on bootstrapping achieves favorable results in varying degrees. The key to achieving high precision and recall is improving the extensibility of patterns. Experimental results demonstrate that the adaptabilities of different seed-selection strategies are varied according to different pattern types, supported by the fact that the extensibility of patterns attains more satisfying results after optimizing the combination of seed-selection and pattern-construction strategies.

This paper uses relational triples generated by ReVerb as the representation of a sentence. The advantage of this strategy is that it contributes to achieving high precision through the reduction of noise during iterations with a relatively simple method. While triple representation contributes to noise reduction, this method also weakens the capability of effective identification of sentences with complex structures. This conforms to our error analysis results, in which there is a considerable proportion of candidate words located in the non-subjective and non-objective parts, such as in objective complement parts. Above all, interpreting a sentence using relational triples constitutes a valid way to identify the core components of a sentence. Nonetheless, it does not fully include a large variety of sentence structures.

This work is therefore far from complete. The proposed method is designed to be domain-independent, i.e. it can be applied to article collections from various research areas. However, We did not verify this with actual data. The approach of triple representation of a sentence is sometimes too simple to contain useful information for DUS identification. We will therefore continue to improve the features used to represent a sentence to better incorporate useful data. Moreover, since our work is still in its early stage and the granularity level of extraction is still too high, more fine-grained extraction will be explored in the future.

Some additional problems can be investigated on the basis of this paper. The first one is the construction of a domain dataset list, which can be constructed by identifying the names of datasets, for which the possible methods can be rule-based filtering or occurrence frequency statistics. The second one is data-driven dataset evaluation. While previous studies on dataset analysis have often used manually collected data, the DUS extraction tool we apply here can be more helpful in evaluating large-scale datasets. In addition, we will explore the use of DUS extraction in scholarly search, and attempt to develop a viable dataset-retrieval service.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No.: 71473183).

Author Contributions

Q.Z. Zhang (zhangqiuzi15@gmail.com) and W. Lu (weilu@whu.edu.cn, corresponding author) proposed the research idea and designed the experiment. Q.Z. Zhang, Y. Huang (hyyc116@gmail.com), and Q.K. Cheng (chengqikai0806@gmail.com) carried out the experiment. W. Lu and Q.Z. Zhang wrote the draft of the manuscript. W. Lu and Q.K. Cheng approved the final version of the manuscript.

References

Aalbersberg, I.J., Dunham, J., & Koers, H. (2013). Connecting scientific articles with research data: New directions in online scholarly publishing. Data Science Journal, 12, WDS235–WDS242.

Belter, C.W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. PLOS One, 9(3), e92590.

Boland, K., Ritze, D., Eckert, K., & Mathiak, B. (2012). Identifying references to datasets in publications. In Zaphiris P., Buchanan G., Rasmussen E., & Loizides F. (Eds.) Theory and Practice of Digital Libraries (pp. 150–161). Heidelberg: Springer.

Chao, T.C. (2011). Disciplinary reach: Investigating the impact of dataset reuse in the earth sciences. Proceedings of the American Society for Information Science and Technology, 48(1), 1–8.

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1535–1545). Association for Computational Linguistics.

Konkiel, S. (2013). Tracking citations and altmetrics for research data: Challenges and opportunities. Bulletin of the American Society for Information Science and Technology, 39(6), 27–32.

Mayernik, M.S. (2013). Bridging data lifecycles: Tracking data use via data citations workshop report. Technical Report, National Center for Atmospheric Research.

Mooney, H., & Newton, M.P. (2012). The anatomy of a data citation: Discovery, reuse, and credit. Journal of Librarianship and Scholarly Communication, 1(1), article no. 1.

Névéol, A., Wilbur, W.J., & Lu, Z. (2011). Extraction of data deposition statements from the literature: A method for automatically tracking research results. Bioinformatics, 27(23), 3306–3312.

Parsons, M.A., Duerr, R., & Minster, J.B. (2010). Data citation and peer review. Eos, Transactions American Geophysical Union, 91(34), 297–298.

Piwowar, H.A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. PLOS One, 6(7), e18657.

Piwowar, H.A., Carlson, J.D., & Vision, T.J. (2011). Beginning to track 1000 datasets from public repositories into the published literature. Proceedings of the American Society for Information Science and Technology, 48(1), 1–4.

Piwowar, H.A., & Chapman, W.W. (2008). Identifying data sharing in biomedical literature. AMIA Annual Symposium Proceedings (pp. 596–600).

Piwowar, H.A., & Vision, T.J. (2013). Data reuse and the open data citation advantage. PeerJ, 1, e175.

Riloff, E. (1996). Automatically generating extraction patterns from untagged text. Proceedings of 13th National Conference on Artificial Intelligence (pp. 1044–1049). Menlo Park, California: AAAI Press.

Robinson, N., Jiménez, E., & Torres, D. (2015). Analyzing data citation practices according to the Data Citation Index. Journal of the Association for Information Science and Technology. http://arxiv. org/abs, 1501.

Thelen, M., & Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10 (pp. 214–221). Association for Computational Linguistics.

Torres, D., Martín, A., & Fuente, E. (2014). Analysis of the coverage of the Data Citation Index–Thomson Reuters: Disciplines, document types and repositories. Revista Española De Documentación Científica, 37(1): 95–97.

This license allows readers to copy, distribute, remix, adapt, and build upon new works non-commercially, and the derivative works are licensed on different terms, provided the original Author(s) and Contribution are credited and the use is non-commercial. Please read the full license for further details at http://creativecommons.org/licenses/by/4.0/