Automatic identification of research articles containing data usage statements*

Qiuzi Zhang

School of Information Management, Wuhan University, NO.
299 Bayi Road, Wuhan, 430072, P.R. China
E-mail: qiuzizhang_whu@foxmail.com

Wei Lu[†], Yunhan Yang
and Haihua Chen

School of Information Management, Wuhan University, NO.
299 Bayi Road, Wuhan, 430072, P.R. China
[†]E-mail: weilu@whu.edu.cn

Jiangping Chen
Information Science, University of North Texas, 1155 Union
Circle #311068
Denton, 76203, U.S.A.
E-mail: Jiangping.Chen@unt.edu

Modern scientific research is characterized with sharing datasets and reusing data for developing new models and theories. This paper describes a study to identify research articles with data use and reuse information. Applying a bootstrapping-based unsupervised training strategy, we were able to develop text patterns automatically out of a large training collection of research articles. These patterns were then used to distinguish articles with data use and reuse from those without data usage. Our experiments using Computer Science literature showed that the identification could achieve more than 85% pattern extensibility. We also demonstrate how the results of the identification could be utilized to gain insights on data sharing and reuse in a scientific field.

Keywords: Data Use and Reuse; Bootstrapping-based Unsupervised Training; Scholarly Text Analysis.

*This work is supported by grant 71173164 of Chinese National Natural Science Foundation.

1. Introduction

Data serve as foundation and evidence for scientific research and progress (Parsons et al., 2010). For instances, test collections developed at TREC (TREC, n.d.) have promoted research and development in information retrieval and access; The ArrayExpress Archive, a repository archiving functional genomics data, has become a major resource for biomedical research community to reuse data for high-throughput functional genomics experiments. Sharing and reuse of scientific data have advanced research and development, boosted knowledge exchange, motivated innovative research design, and accelerated problems solving (Aalbersberg et al., 2013; Chao, 2011; Mooney & Newton, 2012; Piwowar & Chapman, 2008a). Understanding the sharing and reuse of scientific data or datasets is therefore important for data producers, users, and funding agencies. Researchers or funders who make their datasets available would be interested in the citation and impact of their data; Users, especially novice researchers could benefit from data use information to choose and apply appropriate data or datasets in their research.

Even though the importance of data use and reuse has been realized by researchers in different disciplines, the research on data use related issues is unbalanced. Some disciplines, such as Biology, Medicine, and Earth Science, have established several influential data repositories (Robinson-Garciav et al., 2015; Torres-Salinas et al., 2014). Data-related research such as data use tracking (Konkiel, 2013; Mayernik, 2013), analysis of data sharing motivations and impacts (Piwowar, 2011; Piwowar & Chapman, 2008a; Piwowar & Vision, 2013), and datasets evaluation has also been conducted. Researchers in these areas have attempted to incorporate existing data or datasets into their own projects to discover new knowledge. For other disciplines, such as Computer Science and Information Science, there is yet no much investigation on identifying and evaluating data ownership and data reuse.

The purposes of our study were to effectively and efficiently identify articles containing data usage statements, and to explore the application of such identification. Data usage statements (DUS) usually specify how particular data or datasets are obtained, processed, or utilized by authors (Zhang, et al, 2016). These statements allow us to understand data usage behavior reflected in scientific articles. Effective extraction of DUS is therefore crucial, especially in the context of big data, as more and more research articles are available and needed to be analyzed. A few studies have focused on extracting DUS using manual, semi-automatic, or automatic approaches. However, there is no much study on identifying articles containing DUS and applying that to analyze data use and reuse in specific fields.

The rest of the paper is organized as follows: Section 2 reviews existing literature on data usage identification and data usage analysis; Section 3 elaborates the design, procedures, data collections, and experimental setting; Section 4 presents evaluation results; Section 5 discribes an application of using extracted DUS to understand data use behavior in the field of Pattern Recognition; Section 6 discusses the significance and limitations. The paper concludes with a summary and thoughts for future research.

2. Related Studies

This study conducted automatic identification of data use patterns from research articles, and used the identification results to analyze data use and reuse characteristics of a particular field. Thus, we review the related studies in these two areas.

2.1. *Data Usage Identification*

Existing methods on identifying data usage or sharing can be divided into three major categories: (1) human-intensive methods; (2) semi-automatic and machine learning methods; and (3) unsupervised automatic methods.

With human-intensive methods, researchers construct a collection of publications following a certain strategy and then identify whether these publications have used or shared data manually. Piwowar and colleagues (2007) manually examined citation history of 85 cancer microarray clinical trial publications. Some studies retrieved literature from search engines as a candidate collection using Digital Object Identifier (DOI), database accession number, names of data repositories, or other references as queries. Piwowar and colleagues (2011) collected research papers in academic search engines using DOI as search queries. They manually examined whether some specific datasets were used in the retrieved results. Belter (2014) selected three well-known oceanographic datasets to investigate their use and reuse by searching for their names in Web of Science, publisher’s full-text websites and Google Scholar. Some studies were carried out to identify articles in which the authors share their data by providing a link to certain data repositories (Piwowar, 2011; Piwowar & Vision, 2013).

As for semi-automatic methods and machine learning methods, external resources were applied to facilitate the identification of data usage or sharing in some disciplines such as Biomedical science and Geoscience, as these disciplines have established common practice on data sharing and reuse. Piwowar and Chapman (2008a) assumed that an article with data sharing should mention at least the name of a certain data repository. They therefore developed regular expressions containing the names of known data repositories. Different machine learning algorithms were then applied to classify articles using the matched names as features in combination with other features based on bag-of-words. In their following studies (Piwowar, 2011; Piwowar & Vision, 2013), they explored the motivation of data-sharing by judging whether articles about gene microarrays in PubMed shared their datasets. PubMed ID of an article was used as the query when searching the data repositories. Névéol and others (2011) trained a Naïve Bayes classifier and a Support Vector Machine classifier with 586 positive statements and 578 negative statements to extract data deposition statements. The features included the words, the sentence location, the part-of-speech tags, and the sentence composition. Articles containing data deposition behavior were correctly identified with 81% F-measure.

Besides, some studies focused on automatically obtaining data usage in articles. Kafkas and others (2013) studied how database entries were cited in research articles. They conducted the first accession number citation analysis based on the full-text open access articles available from Europe PMC. The BioLit portal provides clickable links from full-text articles to PDB and Gene Ontology based on accession numbers identified in the text (Fink et al., 2008). Haeussler et al., (2011) mined DNA sequences from full text and used them to create links to genomic sequences in Ensembl (Hubbard et al., 2009).

In addition to the methods mentioned above, an automatic and unsupervised strategy called bootstrapping can proceed without external input (Bootstrapping, 2017) , which is a self-starting process. Boland and others (2012) used bootstrapping to identify references to datasets in publications, with the purpose of linking published literatures in social sciences to their corresponding data produced through surveys or interviews. Zhang and colleagues (2016) experimented with a bootstrapping approach to automatically extract data usage statements from academic texts. Our study is a further development from that study (Zhang, 2016), with revisions on the specific steps in the bootstrapping method and DUS discrimination, and extension to analyze data use and reuse characteristics in a scientific field.

2.2. *Data Use and Reuse Analysis*

Open access of scientific data has enabled research on identifying data sharing behavior and its impact on citation and scientific progress (Piwowar et al., 2007; Piwowar and Chapman, 2008b; Piwowar, 2011; Tenopir et al., 2011; Poline et al., 2012; Vines et al., 2014). Most studies concentrated on data reuse behavior, emphasizing the assessment of the value of data reuse (Chao, 2011; Palmer et al., 2011; Meijer et al., 2013; Belter, 2014), the evaluation methodology and measures, and the life cycle and scope of data reuse (Chao, 2011; Piwowar and Vision, 2013).

Faniel and Jacobsen (2010) discovered some key indicators for evaluating reusable data, including data relevance, intelligibility and credibility through interviewing researchers in the field of earthquake engineering. Palmer and others (2011) explored user behavior of data reuse and constructed a data analytic potential model, which consists of potential user communities, preservation convenience, and fitness to purpose. They conducted case studies in Geobiology, Volcanology and Soil Ecology fields to validate the proposed model.

Some researchers have analyzed existing data citation patterns in texts to find out better data citation identification approach. Piwowar and others (2011) selected 1,000 datasets in Biology and earth environment and analyzed the occurrences of their names in articles. They found that data citations were varied without unified, standard formats. Belter (2014) studied the usage of three well-known oceanographic data collections in different tracking methods, finding that citation statistical analysis yielded very different results. Kafkas and others (2013) investigated the citation of some important data repositories in bioscience, finding that citation counts based on mining their accession numbers is as twice as those provided by publishers through structured annotation. In their subsequent research, they extended the scope of their text mining from full text to other supplementary data (Kafkas et al., 2015). Robinson-Garcia and others (2015) analyzed data citation practices based on Data Citation Index from Thomson Reuters and found that there were no shared data citation practices across research fields. While datasets generated in Science and Engineering and Technology were the most cited, those generated in Social Sciences and Arts and Humanities were far less being cited or reused.

Differently, this study analyzed the data use and reuse in the field of Pattern Recognition, a sub-discipline in Computer Science. Applying patterns developed from Computer Science literature, we were able to discover the status and the tendency of data use and/or reuse in Pattern Recognition in combination with traditional bibliometric method.

3. Methodology

This paper proposed and implemented an unsupervised strategy based on bootstrapping to identify data usage at the article level. It determined whether a given article used data, whether built in-house or borrowed from others, to assist research. This section describes the research design, the bootstrapping strategy, experimental design, and evaluation methodology.

3.1. *Research Design*

Figure 1 illustrates the processes of our study, including pattern list acquisition, data usage statements extraction, and article identification/classification based on data usage statements identification. Firstly, some initial seed words as the input of the training process were expanded by bootstrapping to obtain a pattern list through an iterative process. Secondly, the pattern list was employed to extract DUS from an article collection compiled from different sources. Finally, research articles in the collection were classified into one of the two categories: article with data usage vs. article without data usage.

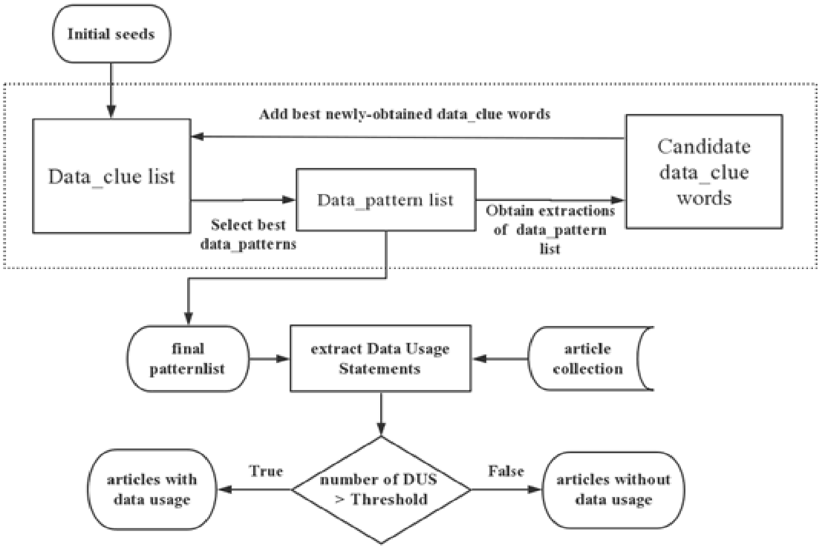

Initial seeds
Add best newly-obtained data_clue words
Data_clue list
Data_pattern list
Candidate data_clue words
Select best data_patterns
Obtain extractions of data_pattern list
final patternlist
extract Data Usage Statements
article collection
articles with data usage
True
number of DUS > Threshold
False
articles without data usage

Fig. 1. Research Design

3.2. *The Bootstrapping Strategy*

We employed a bootstrapping strategy to obtain a pattern list that was used later for DUS extraction. As illustrated in Figure 1 inside the dotted rectangle, the bootstrapping process starts with adding initial seed words as clue words into the data_clue list, called ClueList below (see *Section 3.2.1* for seed words selection). Then perform the following steps (Let’s define the current iteration as *i*, and the maximum of iteration is MAX):

1) Identify and obtain all patterns that match the two pattern types specified in Section 3.2.2 from the relations extracted from a training collection. See section 3.3 for obtaining the relations.
2) Calculate the scores of each pattern P according to Formula 1, and add patterns with top (20+i) scores into the Data_Pattern list.

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

Eq. 1 was first used in (Riloff, 1996) for extraction pattern learning. N refers to the total number of clue words in current ClueList. F refers to the number of clue words contained in this pattern. Each pattern needs to contain at least one clue word, or F>=1. A pattern will be ranked higher if it contains more clue words.

3) Use patterns in the PatternList to extract candidate data_clue words (see pattern examples in Table 1).
4) Calculate score for each candidate clue word with Eq. 2. Add top ranked five new words into the ClueList.

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

Eq. 2 was first used in (Thelen & Riloff, 2002) for semantic lexicons leaning. P refers to number of unique patterns that can extract this candidate word; Fj refers to number of clue words extracted by each pattern. A clue word will be ranked higher if it appears in more patterns.

5) If i < MAX, perform Step 2) to 4). Otherwise, stop the iteration.

The outcome of the bootstrapping process is a list of text patterns.

3.2.1. *Seed words selection*

The seed words that started the bootstrapping training process were manually selected by an inspection of sample papers in our document collection. Three strategies were tested:

1) Selecting the names of a few well-known datasets as seed words. This strategy was abandoned due to poor performance in evaluation experiments;
2) Selecting both names of a few well-known datasets and a few general-purpose words relating to data, such as “dataset” as seed words, which is called COM-SEED for short;

3) Selecting a few general-purpose words related to data, such as "dataset" as seed words, which is called GEN-SEED for short.

Table 3 lists the initial seed words for the training dataset used in this study.

3.2.2. *Pattern Construction*

Pattern in this study refers to a segment of texts presenting a structural feature of a sentence. Patterns can be used to search in texts so matched sentences can be identified and extracted. Linguistically, the more general a pattern, the more sentences can it matches. Considering the degrees of both generalizability and representativeness of patterns, we chose to construct two types of patterns that contain the core of a sentence (the predicate) and the possible positions of the clue words:

1) subject + predicate. In this type of pattern, one or more seed words should appear in the object of the sentence;
2) predicate + object. In this type of pattern, one or more seed words should appear in the subject of the sentence.

Table 1 gives two example patterns and the sentences matching them. The clue words are also highlighted in the sample sentences.

Table 1. Examples of Patterns

Pattern	Matched Sentences with Data Clue Words Highlighted
consist of # samples	**The breast cancer set** consists of 569 samples with 357 benign and 212 malignant
have # samples	Data set 1 is referred to as **Char250**, which has 250 samples per category for lower and upper cases, respectively; data set 2 is referred to as **Char1000**, which has 1000 samples per category for lower and upper cases, respectively. (Note: this pattern occurs twice)
we perform experiment on	To assess the ability of the proposed clustering algorithm to classify the shape classes, we perform experiments on **an increasing number of shapes** in the two Aslan and Tari data sets. We perform our experiments on **a real-estate system** with real-life house dataset used in.

3.3. *The Data Collections for Training and Evaluation*

Table 2 lists the data collections we constructed for this study. Among them, the first five collections are in Computer Science and derived from the ScienceDirect database. The last two collections are in Biomedical Science and derived from a subset of the PMC (PubMed Central) full-text database provided by TREC Clinical Decision Support Track 2016 (Roberts, K, 2016), denoted by PMC OPEN ACCESS. For each data set, an open-course program called ReVerb (Fader, Soderland, & Etzioni, 2011) was employed to extract relations in the triple format of (argument1, relation phrase, argument 2) from each full-text article. Table 2 specifies the size of the relations extracted from each data set. These relations were the input to the bootstrapping process as well as the evaluation experiments.

Table 2. Data Collections Constructed for Training and Testing

Use & Name	Size	Sources
ORIGIN: CSTriples_Whole	39,866,097 relations from 134,610 full-text articles	134,610 full-text articles published between 2000 and 2014 from 115 journals in the field of Computer Science
Train: CSTriples_Train	6,340,339 relations from sections	Sections whose headings contain "result", "experiment" or "evaluation". Selected from 128,317 articles published between 2000 and 2013 of the CSTriples_Whole collection
Evaluation: CSTriples_Test		Articles published in 2014 of the CSTriples_Whole collection
Within-field Evaluation: DATAUSE_INNER	98 full-text articles (84 articles reusing dataset and 14 articles building their own dataset)	Manually annotated articles from CSTriples_Test when they belong to 9 targeted journals (refers to the classification of Artificial Intelligence and Computer Vision and Pattern Recognition in the ScienceDirect database)
Within-field Evaluation: NONUSE_INNER	82 full-text articles without data usage	the same as the data source above

Cross-field Evaluation: DATAUSE_ OUTER	200 full-text articles (randomly select 100 articles from REUSE collection and 100 articles from SELFUSE collection)	REUSE collection is derived from 3355 available articles manually annotate by (Piwowar and Vision, 2013) as reusing the data in GEO repository; SELFUSE collection contains 823 related submission papers crawled from GEO repository
Cross-field Evaluation: NONUSE_O UTER	103 full-text articles	articles citing the submission paper. The citation relations are acquired through querying the PMID of the Submission paper in the Web of Science core collection. The filter criteria require the top 10 citation frequency, and exclude articles with the Accession number detected, and articles with GEO Accession number (such as “GDSnnnn” and “GSEnnnn”) or “data” in the text.

3.4. *The Experiments*

Based on two different seed words selection strategies (COM-SEED and GEN-SEED), we conducted bootstrapping over the CSTriples-Train collection with 300 iterations following the procedures presented in Section 3.2. Table 3 presents the initial seed words used in experiments.

Table 3. Initial Seed Words

Seed selection strategy	Initial seed words
COM-SEED	trec # iris ar face kdd cup uci machine learning repository data/data set/ dataset database corpus
GEN-SEED	data/dataset/data set corpus

With the final pattern list obtained from bootstrapping, we extracted DUS from each paper in the test data collection: subsets of CSTriples_Test and PMC OPEN ACCESS. A sentence in the form of a complete subject-verb-object structure was identified as a data usage statement if it matched at least one pattern in the Data_pattern list.

Finally, each research article in the last four evaluation collections as presented in Table 2 was automatically classified into one of the two categories: articles with data usage vs. articles without data usage, depending on how many DUS it contained. The four evaluation collections were manually annotated beforehand and served as gold standard for this study.

3.5. *Evaluation Measures*

We use *pattern extensibility* to measure the effectiveness of the strategy proposed in this paper for identifying data usage at the article level. In other words, pattern extensibility measures the performance of a computational approach using patterns obtained from a training set to identify data usage from articles in the same or a different subject domain. Specifically, we use *within-field extensibility* to indicate pattern extensibility in the same subject domain, and *cross-field extensibility* to indicate pattern extensibility from articles in a field with a different subject domain.

The evaluation measures for pattern extensibility include Precision, Recall and F measure, which have been used in information retrieval system evaluations and other evaluation tasks. To calculate these measures, we extracted DUS from evaluation datasets with pattern list obtained from training process and constructed respective result datasets. Then, we compared the result datasets with their human annotations. Equations (3), (4), and (5) were used to calculate the scores for precision, recall, and the F measure.

$$\text{Precision} = \frac{Mn}{Rn} \tag{3}$$

$$\text{Recall} = \frac{Mn}{Sn} \tag{4}$$

$$\text{F} - 1 = \frac{2\times\text{Precision}\times\text{Recall}}{Precision+Recall} \tag{5}$$

Here *Rn* denotes the number of all articles in the results collection; *Mn* denotes the number of articles in both results collection and its respective evaluation collection; and *Sn* denotes the number of articles in the evaluation collection.

Among the data collections listed in Table 2, DATAUSE_INNER and NONUSE_INNER were used for *within-field extensibility* evaluation, while DATAUSE_OUTER and NONUSE_ OUTER were for *cross-field extensibility*. To avoid bias caused by the imbalance between the two types of articles, experiments on the DATAUSE_OUTER dataset was repeated for 5 times when evaluating the *cross-field pattern extensibility*, and the average scores were taken as the final evaluation results.

4. Evaluation Results

This section reports bootstrapping and evaluation results on pattern extensibility.

4.1. *Patterns Extracted Through Bootstrapping*

We conducted the iterative bootstrapping processes and extracted clue words and patterns out of CSTriples_Train. Table 4 summarizes the total number of clue words and patterns extracted under different seed word strategies after 300 iterations. In total, we extracted 2,237 patterns using COM-SEED strategy and 1,577 patterns using GEN-SEED strategy.

Table 4. Results of Bootstrapping on CSTriples_Train

Seed word strategy	Pattern Strategy	Number of Clue Words	Number of Patterns
COM-SEED	predicate + object	3462	1007
	subject + predicate	33034	1230
	both	36373	2237
GEN-SEED	predicate + object	3284	360
	subject + predicate	50556	1217
	both	53709	1577

4.2. *Within-field Pattern Extensibility*

We conducted extraction using 4 different combinations of seed word selection strategies and pattern types:

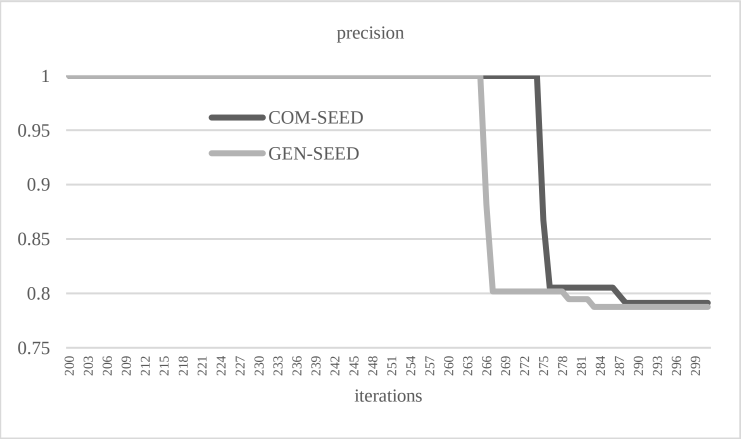
1) COM-SEED. This strategy used COM-SEED to obtain patterns in both pattern types;
2) GEN-SEED. This strategy used GEN-SEED to obtain patterns in both pattern types;
3) COM2PO&GEN2SP, using COM-SEED to acquire patterns of Predicate + Object and GEN-SEED for patterns of Subject + Predicate;
4) GEN2PO&COM2SP, using GEN-SEED to extract patterns of Predicate + Object and COM-SEED for patterns of Subject + Predicate.

Table 5 presents evaluation results on within-field evaluation data collections using the above four pattern acquisition strategies with a threshold of 5 (article contains at least five DUS was identified as data use article). It shows that there are slight differences on identification performance for different pattern strategies. The best performance is achieved under COM-SEED or COM2PO&GEN2SP after 300 iterations. The F-1 value reaches 85.45% after 300 iterations. The COM-SEED strategy produced the best identification performance (F-1 = 86.73% at 275th iterations).

Table 5. Performance of Within-field Data Usage Identification

Pattern Acquisition Strategies	At 300th Iterations			Best Performance During Iterations (based on F-1)		
	Precision	Recall	F-1	Precision	Recall	F-1
COM-SEED	79.13%	92.86%	**85.45%**	86.73%	86.73%	**86.73%**
GEN-SEED	78.76%	90.82%	84.36%	88.17%	83.67%	85.86%
COM2PO&GEN2SP	79.13%	92.86%	**85.45%**	80.53%	92.86%	86.26%
GEN2PO&COM2SP	78.76%	90.82%	84.36%	86.46%	84.69%	85.57%

Figure 2 illustrated the change of the identification performance during 200-300 iterations on precision, recall and F-measure for both COM-SEED and GEN-SEED strategies. It shows that both strategies shared the same changing trends. As the number of iteration increases, the precision of identification decreases while the recall increases until the pattern list acquired being able to cover almost all the evaluation dataset. However, every changing point at the trend line of GEN-SEED strategy appears slightly earlier than COM-SEED, indicating that the speed of pattern acquisition of GEN-SEED is slightly faster.



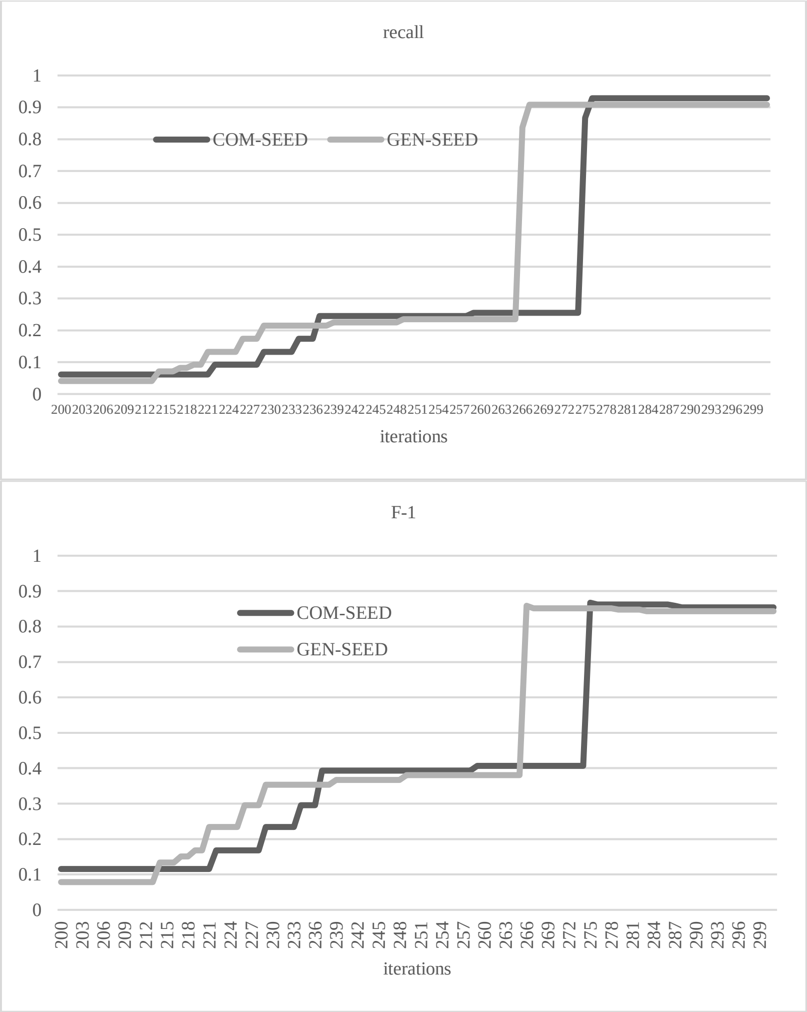


Fig.2. Identification performance over iterations under COM-SEED and GEN-SEED Strategy

4.3. *Cross-field Pattern Extensibility*

As specified in Section 3.3, we used two evaluation data collections in medical domain to test pattern extensibility of extracted patterns. Table 6 presents the cross-field performance with a threshold set to 2. It shows that there is little difference in identification performance among different pattern-acquisition strategies. The F-1 value reached about 88%.

Table 6. Performance of Cross-field Data Usage Identification

Pattern Acquisition Strategies	At 300th Iterations			Best Performance (Based on F-1)		
	Precision	Recall	F-1	Precision	Recall	F-1
COM-SEED	90.95%	85.50%	88.14%	90.95%	85.50%	88.14%
GEN-SEED	90.95%	85.50%	88.14%	90.95%	85.50%	88.14%
COM2PO&GEN2SP	90.94%	85.40%	88.08%	90.94%	85.40%	88.08%
GEN2PO&COM2SP	90.58%	78.50%	86.02%	90.58%	78.50%	86.02%

Figure 3 illustrates the changes of identification performance with increases of iteration under COM-SEED strategy. Table 6 and Figure 3 indicate that the pattern list obtained is independent of subject domains to some extent. In other words, the training results have the possibility to be directly used to different disciplines without repeating the training process.

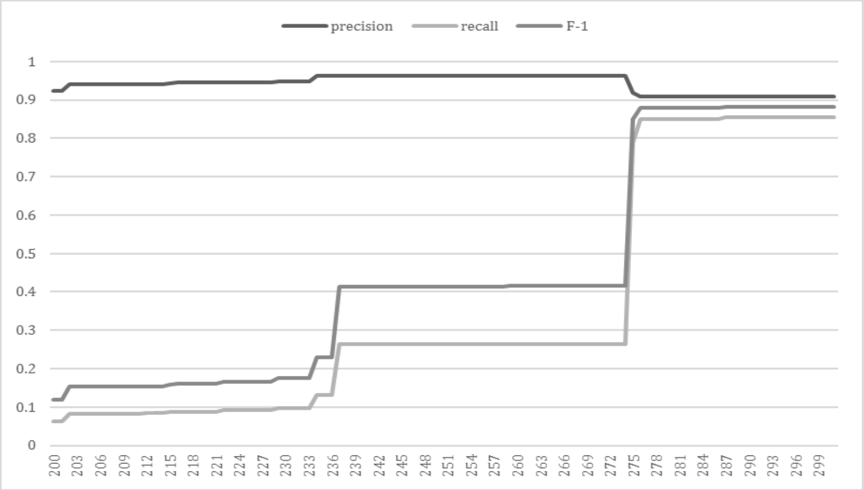


Fig.3. Cross-field Pattern Extensibility under COM-SEED Strategy Over Iterations

5. An Application of Data Usage Identification

The data usage identification method investigated in this paper has many potential applications to facilitate knowledge discovery. To demonstrate this, we carried out an analysis of data usage behavior in Pattern Recognition, a field in Computer Science.

We chose 3,856 articles from the journal called Pattern Recognition for our analysis. The bibliographic data of these articles were obtained through querying the Web of Science using the articles' DOIs. We obtained 3,782 articles out of the 3, 856 with complete bibliographic information including Publication Years, Countries/Territories, Institutions, and Research Areas. We applied COM-SEED strategy with threshold of 5 to classify the 3,782 articles, which results in 2,789 articles with data usage and 993 without data usage.

To facilitate the analysis, we use the term Data Usage Tendency (DUT) to represent the ratio of the number of articles with data usage over the total number of articles sharing the same attribute value. For example, DUT in year 2000 was calculated as the number of articles with data usage published in 2000 divided by the number of articles published in 2000 in the collection of the 3,782 articles.

5.1. *DUT During 2000 – 2014*

Figure 4 demonstrates the change of DUT between 2000 and 2014. It indicates that the number of articles with data usage statements has been increasing year by year during that period. By 2014, the DUT had reached as high as 89%.

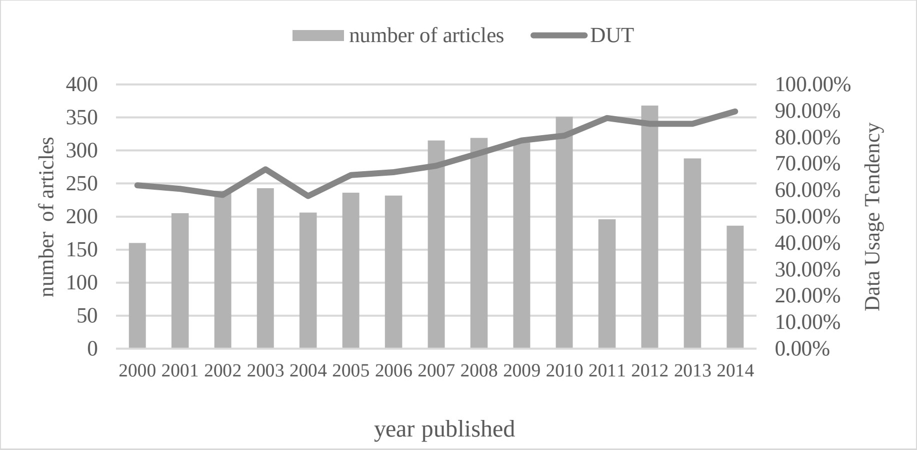


Fig.4. DUT Changes During 2000 - 2014

5.2. *DUT Between Different Countries*

We calculated the number of articles with data usage from 76 countries or regions and present the data usage situation of the top 20 countries in Figure 5. It shows that China, USA, France, UK, and Canada are the top five countries producing most articles containing data usage statements in the field of Pattern Recognition. These countries also published more articles in this field than other countries.

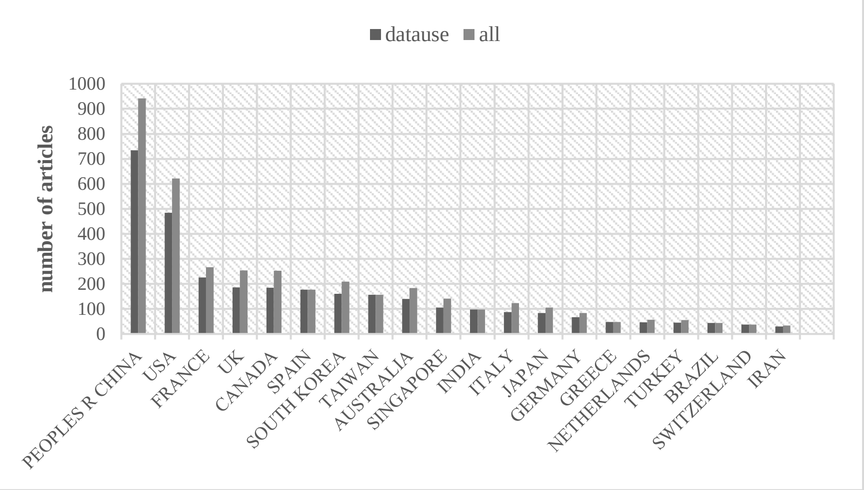


Fig.5. Number of Articles with Data Usage Across Countries or Regions

By examining countries or regions that published more than 50 articles, we found that scores of DUT in high-yielding countries were high. There was no significant difference on DUT scores among those countries, even though they had quite different productivities. Figure 6 illustrates this discovery.

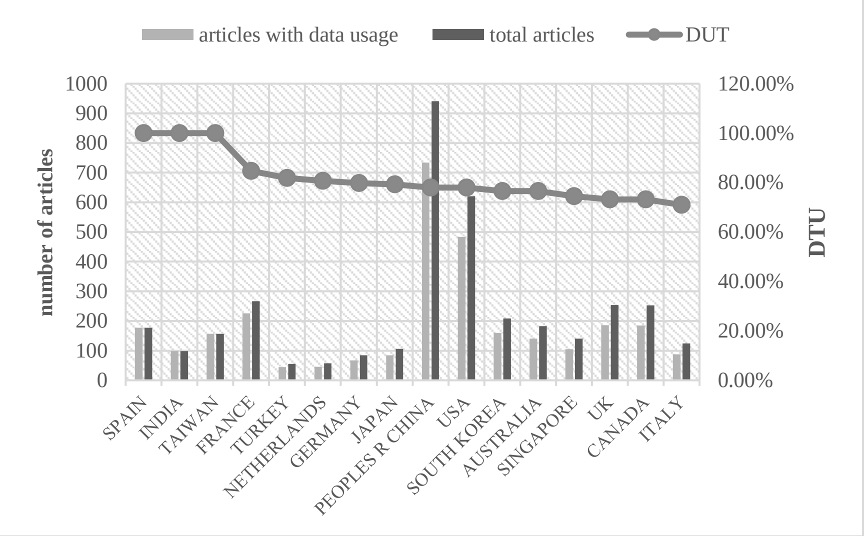


Fig.6. Countries or Regions with Higher DUT Scores

5.3. *Other Discoveries*

We have conducted more analysis, for examples, we identified 32 high-yield institutions and examined their DUT to found out that institutions with high DUT scores usually had a clear data-dependent characteristic; Applying Web of Science Research Area Taxonomy, we identified 15 sub research areas in which the total number of articles was greater than or equal to 50. We found that the data usage tendency of each of these research areas was relatively uniform with very minor difference. Among them, GENETICS HEREDITY and ONCOLOGY, which are research areas closely related to biomedical, are both with higher DUT scores. This finding was consistent with data use and reuse literature indicating that biomedical fields lead data management and sharing among scientific disciplines.

6. Discussion

This study enriched data usage identification and analysis literature by applying a bootstrapping strategy to automatically generate text patterns from a large data collection of Computer Science articles. The bootstrapping strategy enabled the development of a portable pattern list without the need to define relational templates in advance, which had been successfully employed in other studies (Boland et al., 2012; Zhang, et al, 2016). Our study, however, applied the approach differently from previous studies. For example, Boland et al. (2012) used bootstrapping method to linking published literatures in the field of social sciences to the corresponding data produced by questionnaires or interviews. Their judgment on the validity of the pattern was based on setting up a subjective threshold, and the number of initial seed word was confined to only one.

We conducted systematic evaluation of the performance of the bootstrapping strategy. Our overall performance for article-level classification was 85% in terms of F-measure, which is satisfactory as compared to performance reported in the literature using similar approach by Boland and colleagues (2012), which obtained overall 74% F-measure calculated by 97% precision and 60% recall on a different evaluation corpus. Our study approved the effectiveness and efficiency of bootstrapping, especially when there are no well-annotated training materials or templates.

Through the experiments we realized that the threshold of within-field extensibility was larger than that of cross-field extensibility. This indicates that the pattern list obtained from the training set can cover DUS more comprehensively, if the article to be judged is in the same field as training set.

We manually checked the extracted results and found that most patterns in the final pattern list had nothing to do with the discipline, that is, not having unique characteristics of Computer Science. And these patterns played a core role in cross-field extensibility. The final pattern list does contain patterns typical of Computer Science. Therefore, when applying the method to other fields, one can expand training set with articles in the new discipline without much change to the training algorithm. We also demonstrated the use of the classification results in scholarly text analysis and knowledge discovery in this paper.

This study contributes multiple data collections that were constructed in order to implement and evaluate the proposed strategy. These data collections, as listed in Table 2, involved much human effort and can be reused by other researchers.

There are still possibilities to improve our study. Firstly, interpreting a sentence using relational triples is difficult to include a variety of sentence structures. By analyzing the errors in DUS extraction, we found that the pattern obtained in this study might not be sufficient in dealing with the conditions where the seed word appears in the non-subject or non-object position of a sentence. Moreover, not all information in an article had been fully considered for accurately identifying data use and reuse. For example, the information in tables may be able to provide extra help on the performance of the proposed approach.

7. Conclusions and Future Research

This study proposed and implemented a bootstrapping-based unsupervised training strategy to develop text patterns for identifying articles containing data use and reuse statements. Our evaluation experiments showed that the performance of data usage identification at the article level as reflected in its F-measure could be 85% or more, demonstrating that the proposed approach is promising and valuable. The application of data usage identification in Pattern Recognition area facilitated our understanding of the progress and trends of data reuse and sharing in that discipline.

As for the applications of data usage identification and reuse analysis, it will benefit the construction of public datasets. Besides, by identifying data usage, we can build a dataset management system for a specific field to facilitate the dataset sharing among researchers.

Future research includes exploring the relationships between data usage and article elements other than the body of the article, such as article title, footnotes, tables, and charts. Also, we will conduct further examination to understand differences between self-constructed datasets and reuse datasets, and the extraction of data objects that have been reused. Other applications based on data usage identification will also be investigated.

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, September). Identifying references to datasets in publications. In International Conference on Theory and Practice of Digital Libraries (pp. 150-161). Springer Berlin Heidelberg.

Bootstrapping. (2017, April 24). Retrieved May 24,2017, from Wikipedia: https://en.wikipedia.org/wiki/Bootstrapping.

Chao, T. C. (2011). Disciplinary reach: investigating the impact of dataset reuse in the earth sciences. Proceedings of the Association for Information Science & Technology, 48(1), 1-8.

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

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW), 19*(3-4), 355-375.

Fink, J. L., Kushch, S., Williams, P. R., & Bourne, P. E. (2008). BioLit: integrating biological literature with databases. *Nucleic acids research*, 36(suppl 2), W385-W389.

Haeussler, M., Gerner, M., & Bergman, C. M. (2011). Annotating genes and genomes with DNA sequences extracted from biomedical articles. *Bioinformatics*, 27(7), 980-986.

Hubbard, T. J., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., ... & Coates, G. (2008). Ensembl 2009. *Nucleic acids research*, *37*(suppl_1), D690-D697.

Kafkas, Ş., Kim, J. H., & Mcentyre, J. R. (2013). Database citation in full text biomedical articles. *Plos One, 8*(5), e63184.

Kafkas, Ş., Kim, J. H., Pi, X., & Mcentyre, J. R. (2015). Database citation in supplementary data linked to Europe PubMed central full text biomedical articles. *Journal of Biomedical Semantics, 6*(1), 1-7.

Konkiel, S. (2013). Tracking citations and altmetrics for research data: challenges and opportunities. *Bulletin of the Association for Information Science & Technology, 39*(6), 27–32.

Mayernik M S. (2013). Bridging data lifecycles: Tracking data use via data citations workshop report. *NCAR Library*.

Meijer, I., Costas, R., Zahedi, Z., & Wouters, P. (2013). The value of research data - metrics for datasets from a cultural and technical point of view. a knowledge exchange report (april 2013). *J.chem.soc.trans*, *103*, 1774-1789.

Mooney, H., & Newton, M. P. (2012). The anatomy of a data citation: discovery, reuse, and credit. *Journal of Librarianship & Scholarly Communication*, *1*(1), eP1035.

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.

Palmer, C. L., Weber, N. M., & Cragin, M. H. (2012). The analytic potential of scientific data: understanding re-use value. *Proceedings of the American Society for Information Science & Technology, 48*(1), 1-10.

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 Association for Information Science & Technology,48*(1), 1–4.

Piwowar, H., & Chapman, W. W. (2008a). Identifying data sharing in biomedical literature. *AMIA. Annual Symposium proceedings. AMIA Symposium, 2008*, 596.

Piwowar, H. A., & Chapman, W. W. (2008b). Linking database submissions to primary citations with PubMed Central. In *BioLINK Workshop at ISMB*.

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *Plos One*, *2*(3), e308.

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *Peerj, 1*(3), e175.

Poline, J. B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., & Hanke, M., et al. (2012). Data sharing in neuroimaging research. *Frontiers in Neuroinformatics, 6*(6), 9.

Riloff, E. (1996, August). Automatically generating extraction patterns from untagged text. In *Proceedings of the national conference on artificial intelligence* (pp. 1044-1049).

Roberts, K., Demner-Fushman, D., Voorhees, E. M., & Hersh, W. R. (2016). Overview of the TREC 2016 Clinical Decision Support Track. In *TREC*.

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2015). Analyzing data citation practices using the data citation index. Journal of the Association for Information Science and Technology.

Tenopir C, Allard S, Douglass K, et al. (2011). Data sharing by scientists: practices and perceptions. *Plos One, 6*(6), e21101.

Thelen, M., & Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. *Acl-02 Conference on Empirical*

Methods in Natural Language Processing (pp.214-221). Association for Computational Linguistics.

Torres-Salinas, D., Martín-Martín, A., & Fuente-Gutiérrez, E. (2013). An introduction to the coverage of the data citation index (Thomson-Reuters): disciplines, document types and repositories. *Computer Science, 223*(1), 245–250.

TREC mainpage. (2000, August 1). Retrieved May 24,2017, from TREC: http://trec.nist.gov/.

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., ... & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. Current biology, 24(1), 94-97.

Zhang, Q., Cheng, Q., Huang, Y., & Lu, W. (2016). A bootstrapping-based method to automatically identify data-usage statements in publications. *Journal of Data & Information Science*, 1(1), 69-85.