

Identifying and Improving Dataset References in Social Sciences Full-Texts

Behnam Ghavimi ^{a,b}, Philipp Mayr ^a,
Sahar Vahdati ^b, and Christoph Lange ^{b,c}

^aGESIS Leibniz Institute for the Social Sciences,

^bUniversity of Bonn,

^cFraunhofer IAIS

8 June 2016

Outline

- 1 Introduction
- 2 Related Work
- 3 Data Source
- 4 Semi-Automatic Approach
- 5 Evaluation
- 6 Future Work and Conclusion

Motivation

Situations in which explicit links between scientific publications and their underlying datasets are useful :

- Review an evaluation mentioned in a paper.
- Perform further analysis on a dataset used in a paper.

The manual identification of references to datasets in full-text papers

- is time consuming and
- requires hiring some experts in the paper's domain.

Terminology

- A dataset is a **group** of data which are **related** to each other and are the **results** of certain sorts of activities, such as measuring or observing. A data is often prepared due to a scientific research (**purpose**).¹

1. Renear, A. H., Sacchi, S., and Wickett, K. M. (2010)., Definitions of dataset in the scientific and technical literature. DOI: 10.1002/meet.14504701240.

Problem Statement



*Wie erwartet zeigte sich ein Zusammen-
wirtschaftlichen Zukunftserwartun-
aus dem **ALLBUS 2010**, in dem Befragte
Lage einschätzen sollen; vgl. Luszczak*



Dataset Registry



DOI: 10.4232/1.11692

Problem Statement

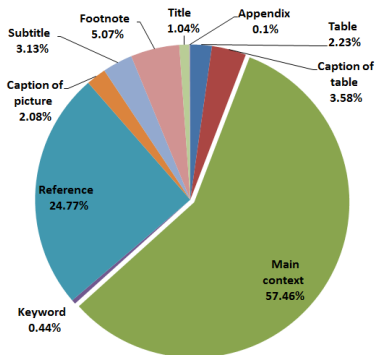


FIGURE – The distribution of dataset references in 15 random mda papers

	Citation style
Paper A	ALLBUS (2010)
Paper B	GESIS – Leibniz-Institute for the Social Sciences: ALLBUS 2010 – German General Social Survey. GESIS, Cologne, Germany, ZA4610 Data File version 1.0.0. (2011–05–30), doi:10.4232/1.10445.
Paper C	ALLBUS (Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften)
Paper D	(e.g., in the German General Social Survey, ALLBUS; see Wasmer, Scholz, Blohm, Walter and Jutz, 2012)
Paper E	Die Einstellungen zu Geschlechterrollen wurden mit Hilfe von Items aus den ALLBUS – Wellen 1994 und 2008 operationalisiert.

FIGURE – Citation styles for a study in five different papers.

Related Work

- Methods Based on the Bag of Words Model
 - Metadata extraction by **tf-idf** [Lee et. al, 2008],
 - Similarity measures such as **Dice2**, **Jaccard**, and **Cosine** [Manning et. al, 1999]
- Corpus and Web Based Methods
 - Dataset extraction by normalized google distance (**NGD**) [Singhal et. al, 2013]
 - Normalized Relevance Distance (**NRD**) [Schaefer et. al, 2014]
- Machine Learning Methods
 - Metadata extraction by using support vector machines (**SVM**) [Zhang et. al, 2006], conditional random field (**CRF**) [Kaur et. al, 2010], and Hidden Markov Model (**HMM**) [Cui et. al, 2010]
 - Dataset extraction by using a bootstrapping approach [Boland et. al, 2012].

Data Source

We use three types of data sources :

- We use full-text articles from the journal **mda**² to evaluate the performance of our approach, and
- Metadata of datasets in the **da|ra**³ registry to identify datasets.(It holds more than **32000** datasets)

2. <http://www.gesis.org/en/publications/journals/mda/>.

3. <http://www.da-ra.de>.

Step 1 : Preparing the Special Features list

- Harvesting DOIs and titles of datasets in da|ra
- Extracting abbreviations and special phrases from Titles.

Examples

- abbreviations :
PIAAC, ALLBUS, EVS, aDvANCE
- special phrases :
Survey of African, Times Poll, People Study

Step 2 : Detecting Dataset References and Ranking Matching Datasets

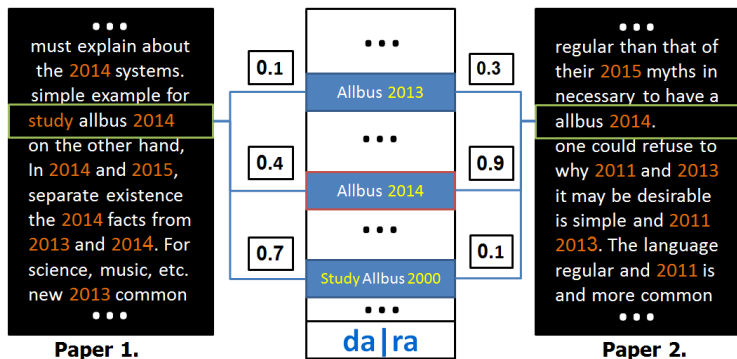
- Detecting the phrases and abbreviations in a paper.
- Comparing **titles** of datasets in da|ra by these **sentences** which include phrases or abbreviations.

Note

Each time a list of titles and a sentence will be compared and both should include a common special features (i.e a common abbreviation or a common special phrase).

Applying **Tf-idf** on list of titles and the sentence for vectorizing them and then **cosine similarity** give a similarity score to each title on the list based on the sentence.

Step 3 : Heuristics to Improve Ranking in Step 2



Step 4 : Exposing the Results to the User, and Interactive Disambiguation

Two main work flows :

1 By reference :

for each reference, five titles of candidate datasets are suggested to the user. While this workflow best supports the user in getting every reference right, it can be tiring.

2 By characteristic feature :

It suggests six titles of candidate datasets to the user for each feature (which may be common to multiple individual references in the paper).

An overview of the approach

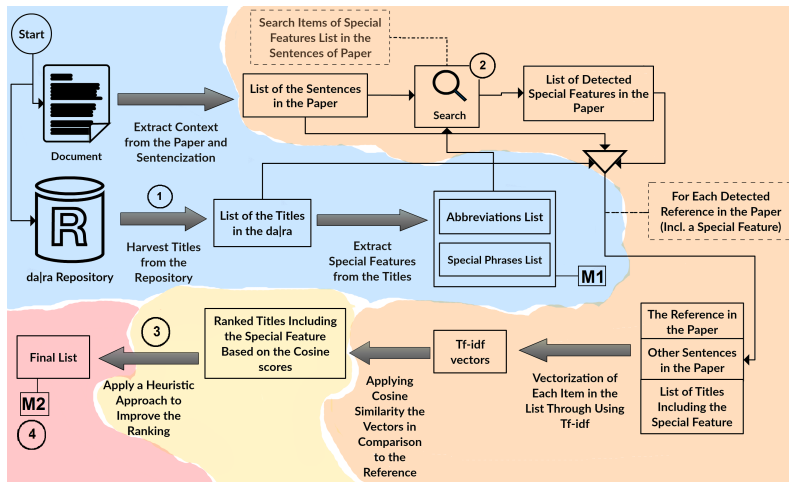


FIGURE – An overview of the approach.

Evaluation Test Corpus and Gold Standards

- Our test corpus included **15 random papers** from the **2013** and **2014** issues of the mda journal – **6 in English** and **9 in German**.
- A trained assessor from the InFoLiS II⁴ project at GESIS reviewed all papers one by one and **identified all references to datasets**.
- Afterwards, the assessor attempted to discover a correct match in da|ra for each detected reference, resulting in a list(**gold standard**) of datasets per paper.

An Example for Evaluation

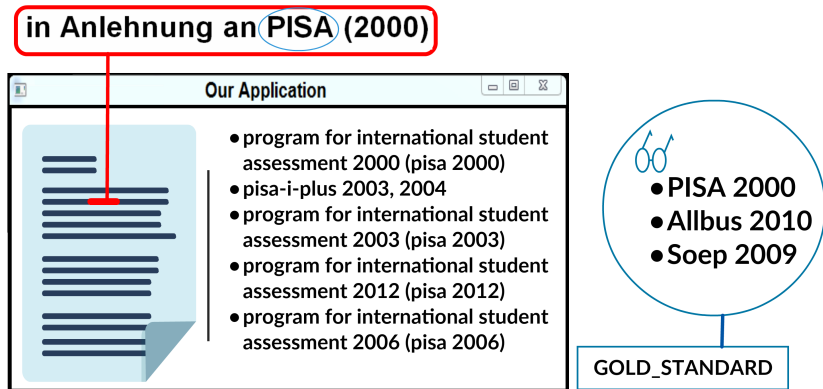


FIGURE – The example illustrates the evaluation section.

Evaluation Results

- More Information about Test Corpus (without considering **different versions** of each dataset, **25** datasets exists in our test corpus) :

	Max.	Min.	Avg.
Datasets in a Paper	7	1	4
References to a Dataset	147	1	12

- Results of the Evaluation :

Phase of Evaluation	Precision	Recall	F-measure
Detection	0.91	0.77	0.84
Matching	0.83	0.83	0.83

Conclusion and Future Work

Conclusion :

- no cold start problem.
- returns a ranked list.
- F-measure about 0.83.
- generates results very fast.

Future work will focus on improving **accuracy** and on extending **coverage** :

- taking **synonyms** into account.
- identifying **central dataset(s)**.
- taking **further datasets** not registered in da|ra into account.
- using a combination of the titles and **other metadata**.





Lee, S. and Kim, H.J., 2008, September. News keyword extraction for topic tracking. In Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on (Vol. 2, pp. 554-559). IEEE.



Manning, C.D. and Schütze, H., 1999. Foundations of statistical natural language processing (Vol. 999). Cambridge : MIT press.



Singhal, A. and Srivastava, J., 2013. Data extract : Mining context from the web for dataset extraction. International Journal of Machine Learning and Computing, 3(2), p.219.



Schaefer, C., Hienert, D. and Gottron, T., 2014. Normalized Relevance Distance-A Stable Metric for Computing Semantic Relatedness over Reference Corpora. In ECAI (Vol. 263).



Zhang, K., Xu, H., Tang, J. and Li, J., 2006. Keyword extraction using support vector machine. In Advances in Web-Age Information Management (pp. 85-96). Springer Berlin Heidelberg.



Kaur, J. and Gupta, V., 2010. Effective approaches for extraction of keywords. Journal of Computer Science, 7(6), pp.144-148. Vancouver



Cui, B.G. and Chen, X., 2010. An improved hidden Markov model for literature metadata extraction. In Advanced Intelligent Computing Theories and Applications (pp. 205-212). Springer Berlin Heidelberg.



Boland, K., Ritze, D., Eckert, K. and Mathiak, B., 2012. Identifying references to datasets in publications. In Theory and Practice of Digital Libraries (pp. 150-161). Springer Berlin Heidelberg.