Identifying and Improving Dataset References in Social Sciences Full–Texts

Presenter:

Behnam Ghavimi

Supervisor:

Prof. Dr. Sören Auer
Dr. Philipp Mayr

15 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

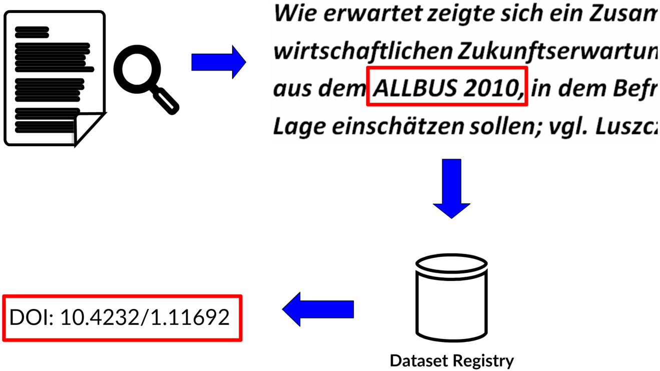
- is time consuming and
- requires hiring some experts in the paper's domain.

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 Zusan
wirtschaftlichen Zukunftserwartun
aus dem ALLBUS 2010, in dem Befr
Lage einschätzen sollen; vgl. Luszc

Problem Statement

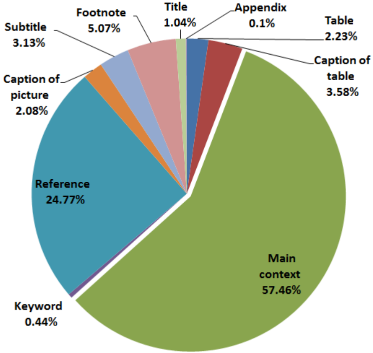
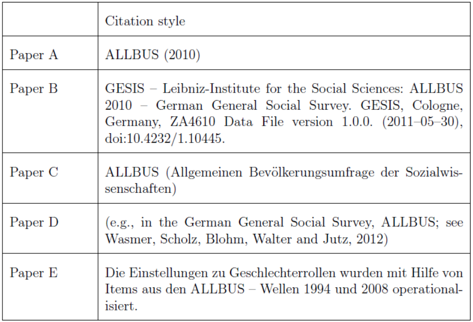


FIGURE – The distribution of dataset references in 15 random mda papers



	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)
- Metadata of papers in the **SSOAR**[4] repository for exporting the suggestions of our approach for papers as **JSON** files.

[2] **http://www.gesis.org/en/publications/journals/mda/**.

[3] **http://www.da-ra.de**.

[4] **http://www.ssoar.info**.

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**, **Singularisierungsstudie**

Step 2 : Detecting Dataset References and Ranking Matching Datasets

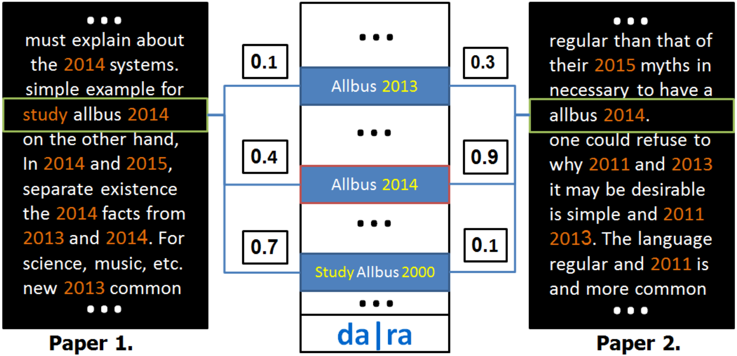
- Detecting the phrases and abbreviations in a paper.
- Comparing **titles** of datasets in da|ra by these **sentences** which include phrases or abbreviations.

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



Paper 1.

• • •
must explain about
the 2014 systems.
simple example for
study allbus 2014
on the other hand,
In 2014 and 2015,
separate existence
the 2014 facts from
2013 and 2014. For
science, music, etc.
new 2013 common
• • •

Paper 2.

• • •
regular than that of
their 2015 myths in
necessary to have a
allbus 2014.
one could refuse to
why 2011 and 2013
it may be desirable
is simple and 2011
2013. The language
regular and 2011 is
and more common
• • •

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

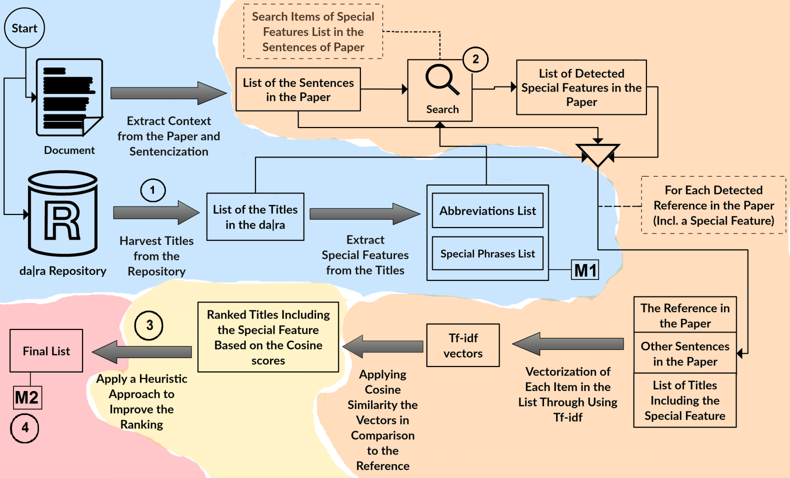

Start
Document
Extract Context from the Paper and Sentencization
Search Items of Special Features List in the Sentences of Paper
List of the Sentences in the Paper
Search
2
List of Detected Special Features in the Paper
da|ra Repository
1
Harvest Titles from the Repository
List of the Titles in the da|ra
Extract Special Features from the Titles
Abbreviations List
Special Phrases List
M1
For Each Detected Reference in the Paper (Incl. a Special Feature)
The Reference in the Paper
Other Sentences in the Paper
List of Titles Including the Special Feature
Vectorization of Each Item in the List Through Using Tf-idf
Tf-idf vectors
Applying Cosine Similarity the Vectors in Comparison to the Reference
Ranked Titles Including the Special Feature Based on the Cosine scores
3
Apply a Heuristic Approach to Improve the Ranking
Final List
M2
4

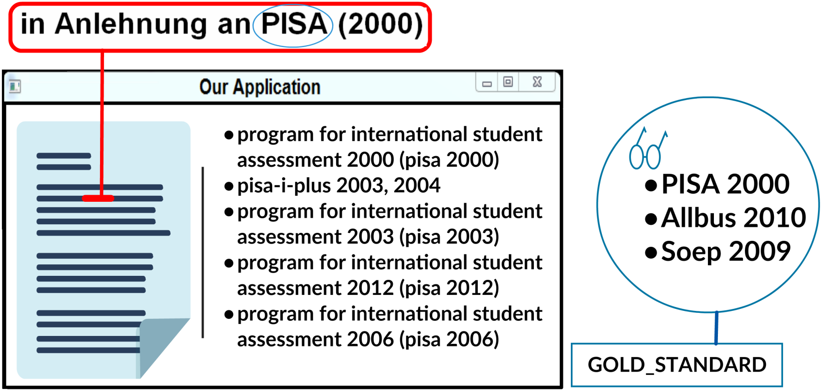
FIGURE – An overview of the approach.

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 II[5] 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.

[5] **http://www.gesis.org/forschung/drittmittelprojekte/projektuebersicht-drittmittel/infolis-ii/**.

An Example for Evaluation



in Anlehnung an PISA (2000)

Our Application

- program for international student assessment 2000 (pisa 2000)
- pisa-i-plus 2003, 2004
- program for international student assessment 2003 (pisa 2003)
- program for international student assessment 2012 (pisa 2012)
- program for international student assessment 2006 (pisa 2006)

- PISA 2000
- Allbus 2010
- Soep 2009

GOLD_STANDARD

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
Detection+Matching	0.76	0.64	0.7

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.

Preliminaries

Tf-idf (applied on **all sentences** and **titles**) :

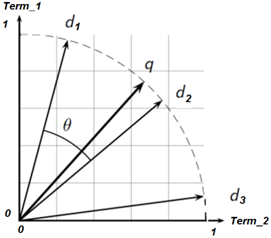
$$w_{t,d} = \begin{cases} 1 + log_{10} tf_{t,d}, & \text{if } tf_{t,d} > 0 \\ 0, & \text{if } tf_{t,d} = 0 \end{cases}$$

$$idf = \log_{10}(N/df_t)$$

$$\text{tf-idf (q,d)} = \sum_{t \in q \cap d} tf.idf_{t,d}$$

Cosine-Similarity (applied on **a dataset reference** and **titles**) :

$$\cos(\vec{q}, \vec{d}) = \cos\theta = \frac{\vec{q} \cdot \vec{d}}{\| \vec{q} \| \| \vec{d} \|} = \frac{\vec{q}}{\| \vec{q} \|} \cdot \frac{\vec{d}}{\| \vec{d} \|}$$



Preliminaries

Other Similarity Algorithms :

- □ Matching Coefficient=$|d \cap q|$
- □ Dice Coefficient=$\frac{2|d \cap q|}{|d|+|q|}$
- □ Overlap Coefficient=$\frac{|d \cap q|}{min(|d|,|q|)}$
- □ Jaccard Coefficient=$\frac{|d \cap q|}{|d \cup q|}$

Preliminaries

Evaluation Metrics :

$$\text{Precision}=\frac{\#\text{True Positives}}{\#\text{True Positives}+\#\text{False Positives}}$$

$$\text{Recall}=\frac{\#\text{True Positives}}{\#\text{True Positives}+\#\text{False Negatives}}$$

$$\text{F-measure}=2\cdot\frac{\mathit{Precision}\cdot\mathit{Recall}}{\mathit{Precision}+\mathit{Recall}}$$

Note :

-**High** **Precision** shows **less** **FP**.
-**High** **Recall** shows **less** **FN**.
- **F-measure** is the **harmonic mean** of precision and recall.

JSON files' properties (Source : Schema.org)

Properties from Creative Work

Property	Expected Type	Description
headline	Text	Title of the item.
alternativeHeadline	Text	A secondary title of the item.
author	Person or Organization	The author of this content.
citation	Creative Work or Text	A reference to another creative work, such as another publication, dataset, etc.
inLanguage	Text or Language	The language of the content.
datePublished	Date	Date of first publication.
publisher	Person or Organization	The publisher of the item.

Properties from Thing

Property	Expected Type	Description
description	Text	A short description of the item.
sameAs	URL	A URL to address the items on the web explicitly.

Positioning and Power in Academic Publishing: Players, Agents and Agendas
Proceedings of the 20th International Conference on Electronic Publishing
Edited by
Fernando Loizides
Birgit Schmidt
ELPUB 2016
IOS Press

- International Conference on Electronic Publishing (ELPUB) - ELPUB 2016, 7–9 June 2016 in Göttingen, Germany
- Ebook : Positioning and Power in Academic Publishing : Players, Agents and Agendascover Proceedings of the 20th International Conference on Electronic Publishing ,Published in 2016, Editors : Fernando Loizides, Birgit Schmidt
  ISBN :978-1-61499-648-4 (print) and 978-1-61499-649-1 (online)

ELPUB 2016 Positioning and Power in Academic Publishing: Players, Agents and Agendas

IOS Press