Towards a Similarity Algorithm for Controlled Vocabularies within the Digital Humanities

Felix Ernst[0000−0002−2102−4170]

Karlsruhe Institute of Technology, Karlsruhe, Germany
felix.ernst@kit.edu

Abstract. With a growing amount and increasing complexity of data and metadata in the Digital Humanities, the use of semantic tools such as controlled vocabularies and taxonomies becomes more and more important to gain new research insights. Their use enables new research possibilities by introducing machine readable semantic links and standardised data and metadata. A validation and recommender system that ensures a quick development of high quality vocabularies is essential in such a scientific workflow. The base of this system is a similarity algorithm. State of the art algorithms and editors for controlled vocabularies do not meet the special requirements of the Digital Humanities domain. Therefore, this work proposes to fill the research gap in the Digital Humanities domain with a similarity algorithm and a recommender and validation system for controlled vocabularies. The methodology and evaluation for achieving this goal as well as preliminary results are presented in this contribution.

Keywords: Semantic Web · Similarity metrics · Recommender system · Controlled vocabularies · Vocabulary editor · SKOS

1 Introduction and Motivation

Computer-based methods for answering research questions in the Digital Humanities (DH) pose special challenges: The research data are diverse and oftentimes do not consist of machine readable text but rather of textual fragments, images, 3D models, illustrations and many more, all in multiple (historical) languages and writing systems. Additionally, they are often incomplete, distributed across different data sources or over multiple countries and growing in complexity. Knowledge enrichment of data and metadata by semantic methods plays an important role to overcome these difficulties [6].

This can be best illustrated on a simplified exemplary DH use case of an ongoing research project that is studying different historical language learning books which is part of the Collaborative Research Centre (CRC) 980 'Episteme in Motion'[1]. The group's research interests focus on various aspects of the textbooks, one of which is the book's target audiences. The scholars annotate the

[1] funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), see https://www.dfg.de/sfb/.

digital images of the book pages with machine readable tags to enable further data analysis. A problem arises when tagged historical terms that refer to the same target group differ due to grammar, synonyms or different languages. A controlled vocabulary[2] for all tags solves this problem by semantic links. Synonyms or terms in different languages are incorporated and linked to the single item they refer to. Together with a vocabulary for descriptive metadata about the books such as author, place of printing or writing language, it results in a Knowledge Organisation System (KOS) with well categorised and interlinked tags and metadata. This enables the development of advanced computer-assisted methods for analysing, managing and visualising data. In the DH case, existing vocabularies found in vocabulary registries oftentimes are either too broad or do not match the scholar's needs or required language. Thus, there is a need for developing own vocabularies that are tailored to the data, to the research questions and that are easily shareable. This gives rise to new research possibilities since the reuse of results as well as the collaboration of both disciplinary and interdisciplinary research groups is simplified [5]. Unfortunately, building subject-specific, high-quality vocabularies often is not feasible for domain experts in a reasonable amount of time with currently available vocabulary editors. This is due to the fact that these are either outdated, lack well-written documentation and usability or do not support widely used standards which prevents their use.

One important step to build vocabularies in a fast and efficient way is the use of validation and recommender systems. The latter presents terms and term sets from existing vocabularies and knowledge bases to the researcher that match the current topic or field. Integration of the recommendations allows for a faster vocabulary development. A semantic and content-wise validation system ensures accurate and error-free high quality vocabularies [15]. Both systems rely on similarity algorithms for vocabularies and its terms. The algorithms are not only capable of finding similar terms but also compare semantic relations. Developing a semantic similarity algorithm and a validation and recommender system for the DH faces several challenges. The number of relevant existing domain specific vocabularies or knowledge bases is oftentimes low. Additionally, not only the research data are multilingual, but also the research and the output itself which results in challenges when mapping terms of different (sometimes dead) languages. Furthermore, material and human resources of DH projects often are tight. This leads to the need of specifically designed similarity metrics.

These and the validation and recommender systems have to meet the following **requirements** in order to overcome the aforementioned challenges of the DH: (a) support of heterogeneous, multilingual vocabularies and data; (b) applicability even when preexisting vocabularies are scarce; (c) high usability since DH scholars typically have less background in technical computer science issues such as algorithm improvement or adapting research software; (d) utilisation in projects with limited human resources for acquiring knowledge about ontology development; (e) applicability in projects with restricted possibilities of running resource intensive algorithms or software.

[2] In the following, thesauri and taxonomies are subsumed under the term vocabulary.

The overarching goal is to propose a validation and recommender system for vocabularies which can easily be integrated into a vocabulary editor. It aims to support researchers by lowering the barrier of entry for building high-quality vocabularies. No prior knowledge is needed, thus it facilitates computer-aided research.

2 State of the Art

In order to find relevant vocabularies or terms for a particular research area, there are various **vocabulary registries** that contain humanities specific terms. These include *Linked Open Vocabularies*[3], *DARIAH EU back bone thesaurus*[4], *Linked open terminology resources*[5], *CLARIAH awesome humanities ontologies*[6] or the *Basic Register of Thesauri, Ontologies & Classifications*[7]. Some exemplary tasks taken from sub-projects of CRC 980 include the description of old Greek texts and Egyptian hieroglyphs as well as multilingual terms for ancient plant names. For these tasks, no adequate vocabularies could be found due to the lack or incompleteness of non-English vocabularies. Nevertheless, registries are a valuable source to find vocabularies that can serve as starting point for the development process.

SKOS[8] (Simple Knowledge Organization System) is a data model for representing controlled vocabularies which became a formal W3C recommendation in 2009. It is well suited for building vocabularies because there are several advantages compared to other data models such as a simple integration in the Semantic Web, a flexible and standardised development and its simplicity [10].

Having a suitable data model for controlled vocabularies is not sufficient. A **vocabulary editor** that supports the researcher during the entire process and facilitates building, curating and publishing SKOS vocabularies is essential. In addition to the special DH requirements mentioned in section 1, the application in research projects poses additional ones such as open source software, a customisable web interface, exchangeable backend storage, importing vocabularies as well as a flexible user management. These are taken into account but are not the scope of this contribution and thus not elaborated on further. Although numerous tools were examined (40 in total including Protegé, CESSDA Vocabulary Editor, Neologism 2.0, Vocbench, iQVoc, Bioportal, Vocoreg, Themas, HIVE, NERC Vocabulary Server, OpenSKOS, PoolParty, SissVoc, TemaTres, Unilexicon, VocPrez, Wikibase), none of these fulfilled the requirements or were easily extendible to do so. Furthermore, there is always a trade-off between usability and the number of features. For instance, Protegé and VocBench are feature rich and in general suitable for a large amount of classes and triples. When only

[3] https://lov.linkeddata.es/
[4] https://www.backbonethesaurus.eu/
[5] https://www.loterre.fr/
[6] https://github.com/CLARIAH/awesome-humanities-ontologies/
[7] https://bartoc.org/
[8] https://www.w3.org/TR/skos-reference/

using a small subset of the functions to build a SKOS vocabulary, both tools are not able to support and guide a researcher well in the development process who is typically not an ontology expert. This leads to a high barrier in adopting the tools into the daily routine. The Austrian Centre for Digital Humanities (ACDH) at the Austrian Academy of Sciences is developing an editor within the DH context.[9] At the time of evaluation, it was still in the development process and could not be assessed in regards to all key requirements.

Similarity algorithms for vocabularies are able to quantify the similarity of two terms or two term sets with an individual semantic structure. They form the basis for a recommender system and a content-wise validation for vocabularies. The similarity measures can be split into two different main methods: Deep learning and non deep learning methods. In the former, different measures using neural networks are used to compute similarity between texts, phrases or terms which requires a large number of domain specific vocabularies [9]. Those exist in disciplines such as biomedical sciences, but usually not in the humanities or small disciplines. Furthermore, the infrastructure requirements for deep learning methods are demanding, as stated by Nguyen et al. [9], and often cannot be met by a large part of research projects, especially in the humanities and smaller fields or projects. Since preparing a well suited training dataset and performing algorithm training is not feasible for most projects (lack of human, material and fitting data resources), it would only be possible to provide a training model for similarity prediction. However, this would introduce a quality loss when applied to other languages or fields. This contradicts the aim for a sustainable algorithm that can be applied independent of domain and language without the need of constant improvement. For an overview about deep learning methods see [2].

Non deep learning methods can be further broken down into text corpus-based[10] methods and knowledge-based methods. Similarity measures using a text corpus are used to thematically group terms and phrases and generate a vocabulary [2]. These are based on the 'distributional hypothesis' that 'similar words appear in similar contexts' [3]. Following the distributional model, a large corpus is needed such that infrequent words can be represented accordingly. For the previously presented DH use case, such a large amount of data, especially machine readable digitised texts, is not available. Hence, corpus-based methods are not suitable in this case.

Knowledge-based methods (KB methods) use sources with structured knowledge content such as ontologies, thesauri or lexicons for defining the similarity of terms or vocabularies [2,4]. The different KB methods are presented and contextualised in the following. The simplest way to define similarity is to take the taxonomical structure into account and calculate the path length between two terms in a KB tree as proposed by Rada et al. [11]. The lower the distance, the closer the relationship between two terms. Path length based methods rely on well built knowledge bases by domain experts that contain all relevant terms. If a term is missing, the path length and thus the similarity cannot be calculated.

[9] https://github.com/acdh-oeaw/vocabseditor/

[10] A text corpus is a large, structured collection of texts.

The overall goal of the proposed algorithm is to eventually support domain experts in building high quality vocabularies. Such vocabularies would be needed for all path based measures but these do not exist yet. Hence, this approach is not feasible for our case.

A more advanced KB method is based on the information content (IC) of a term. IC can be described as 'the amount of information provided by the term when appearing in a context' [13]. The IC is either derived from the Inverse Document Frequency of the term in a text corpus [2] or from the structure of the knowledge base itself [14]. The former depends on text corpora, the latter on a pre-existing, distinct and well built vocabulary. Since neither is available in the DH use case, IC based methods are not suitable in this case.

Another KB method compares the features and attributes of two terms [4]. The similarity increases the more features and attributes they share, such as description, related terms and others [13,2]. In particular, the overlap of term descriptions can be well suited to define semantic similarity [1]. This approach is suitable for the present use case because multiple vocabularies can be incorporated and used for similarity computation, a pre-existing universal vocabulary is not needed. In case that a term lacks attributes but includes information about exact or close matches in a network KB such as Wikipedia[11], this can be exploited to calculate similarity using the description and semantics of the term in the network knowledge-base [7].

Apart from a plain syntax validation, two types of **vocabulary validation** are introduced in the following. Semantic validation means that there are no logical errors and there is no violation of the data model. Skosify is a Python library which provides such a validation [15] and is well suited for the integration into the proposed validation system. Content-wise validation means that a vocabulary which follows all SKOS rules does not include content that would be judged wrong by a user, e.g. assigning a term accidentally to the wrong branch in the hierarchy. Furthermore, vocabularies such as the Shapes Constraint Language[12], Shape Expressions[13] or Resource Shape[14] are promising candidates for specifying integrity constraints in the validation system. Up-to-date, there exists no tool that provides content-wise validation for SKOS vocabularies which poses a research gap that is addressed in this work.

Concerning **recommender systems** for vocabularies, only Neologism 2.0 [8] offers a basic one. The search of its recommender system is mainly based on term labels which means that semantic features are omitted when giving recommendations. This leads to a limited value because only terms with an identical label in external vocabularies can be found. Synonyms, different languages or a different spelling for relevant terms prevent the recommender of finding them. Hence it falls short of the potential of recommenders for vocabulary development.

[11] https://en.wikipedia.org/
[12] https://www.w3.org/TR/shacl/
[13] https://www.w3.org/community/shex/
[14] https://www.w3.org/Submission/shapes/

3 Problem Statement and Contributions

Hypothesis: The developed similarity metrics deliver better results than the state of the art concerning suitability in small research fields, resource consumption and application to a multilingual database.
RQ1: To what extent are knowledge-based similarity algorithms superior to other methods when calculating the similarity of controlled vocabularies in the DH domain?
RQ2: To what extent can a knowledge-based similarity algorithm be modelled to be applicable for small research fields, low resource consumption and a multilingual database as found in the DH context?
RQ3: To what extent can the resulting algorithm be applied to other disciplines, e.g. materials science?
RQ4: To what extent can the recommender and validation system support the development of subject-specific, multilingual DH vocabularies?

The research will contribute as follows:

- Design of a reference vocabulary in the field of DH for evaluation of similarity algorithms,
- development of a similarity algorithm for vocabularies suitable for small research fields, low resource consumption and a multilingual database,
- elaboration of a semantic and content-related validation for vocabularies,
- design of a recommender system for vocabularies,
- evaluation of the recommender and validation system within multidisciplinary projects of the CRC 1475 'Metaphors in Religion' and the CRC 980 'Episteme in Motion'.

4 Research Methodology and Approach

In this section, the methodology and approach is elaborated for each research question.

RQ1: The first step is to collect state of the art similarity metrics that are in general suitable for vocabularies. All gathered methods are then evaluated with respect to the present DH use case and its specific demands. As a result, the best state of the art algorithms are assessed and their performance quantified. This is done using a DH reference vocabulary whose design is also part of the research since there are no such reference datasets available.

RQ2: So far, there exists no knowledge-based similarity algorithm that fits the requirements of the present DH use case. Therefore, the challenge is to develop an algorithm that is capable of providing sufficient results in similarity detection while being used in small research fields with a limited number of knowledge bases and vocabularies, low resource consumption and suitable for a multilingual database. It needs to be assessed to what extent existing algorithms can serve as a starting point for the algorithm development.

RQ3: The objective is to provide a similarity algorithm that is not limited to DH but adaptable to other disciplines, for instance material science. Hence, close cooperation with researchers of other domains is established such that different reference vocabularies can be provided and evaluated. The results are then compared to the DH use case and to other algorithms.

RQ4: The recommender system proposes matching terms or branches of external vocabularies during the development process. The validation system evaluates if there are any semantic or content-wise mistakes by comparing neighbouring terms in both the source and the external vocabulary and by using predefined integrity constraints. If an internal threshold is reached which suggests a content-wise mistake, there will be a corresponding user output. Both systems will be integrated into a vocabulary editor which is currently being developed within the information infrastructure sub-project of CRC 980. Close cooperation is already established to researchers of CRCs 980 and 1475 which allows for receiving early user feedback and eventually answer the research question.

5 Evaluation Plan

5.1 Reference Datasets and Algorithm Evaluation

A crucial element in designing similarity metrics is performance monitoring including a comparison to the state of the art as early as possible. To achieve this, a domain specific vocabulary is built together with DH scholars and serves as base data for evaluation. Since there is no default way to objectively rate the accuracy of computational similarity, the results for each algorithm can be compared to human similarity ratings given by domain experts which represent the baseline [12]. Even though machine learning based methods were shown to be unsuitable for the presented case, their results are also compared to the developed similarity metrics.

Furthermore, the computational similarity of terms in the domain specific vocabulary and other terms in multiple publicly available vocabularies is used as performance indicator. To obtain domain independent results, vocabularies outside the DH are considered as well. If it is not feasible to obtain human similarity judgements as baseline for domain independent vocabularies, high quality lexical-semantic networks such as WordNet[15] or GermaNet[16] are used. In this case, the focus is to find close or exact matches of terms (meaning high computational similarity) and compare it to the ground truth (meaning synonyms) as specified by the utilised lexical-semantic networks.

5.2 Validation and Recommender System Evaluation

The semantic and content-wise validation of vocabularies is evaluated by randomly modifying the vocabulary created by domain experts and introducing

[15] https://wordnet.princeton.edu/
[16] https://weblicht.sfs.uni-tuebingen.de/rover/

false content and wrong semantic links, e.g. closed loops. To ensure neutrality, the modification is done by independent individuals. When performing the validation, the number of detected faults or imperfections in the vocabulary can be quantified and compared to the actual number of introduced errors.

The performance assessment of the recommender system is challenging. Since it is highly subjective if a recommendation is helpful or not, competency questions are defined together with domain experts. This means that fragments of vocabularies X are given as system input ('Which terms and/or vocabularies are similar to X?'), the output ('Term t and vocabulary Y have high computational similarity to X') is then compared to what domain experts are expecting or considering as helpful. The question formulation is done in close contact with researchers of ongoing DH projects within the CRCs 1475 and 980. To avoid tuning the algorithm's performance to the competency questions, these are formulated on an ongoing basis during the whole development process.

6 Preliminary Results

Since this work is at an early stage, only preliminary results regarding similarity algorithms and vocabulary editors are presented. To find state of the art similarity algorithms that are well suited for the DH case, a literature study was conducted. The algorithms were classified into different groups and evaluated. The results are the basis of this work and are summarised in section 2.

Concerning the vocabulary editor, a survey with prospective users of four different humanities projects was conducted to determine the needs of the user base. The state of the art was evaluated against these requirements and is planned to be published as survey paper because to the best knowledge this has not been done so far. As the currently available methods do not fulfil the requirements, a basic vocabulary editor was developed in cooperation with computer science students. This editor addresses the additional, domain independent requirements outlined in section 2: It is written in python, easily extensible, uses SKOS as data model, offers a web interface, provides user management and is capable of collaboratively developing vocabularies.

To include future users as early as possible on in the development process, a hands-on workshop for scholars was held in October 2021 where the participants developed a simple vocabulary and used it to annotate digital images which closely resembles the present DH use case. Additionally, the editor and its prospective use was presented in September 2021 to members of the German engineering community to include future fields of application early on. As a next step, the DH reference vocabulary will be addressed and built such that algorithm performance can be quantified.

7 Conclusions and Lessons Learned

This work contributes to the design, development and evaluation of a validation and recommender system for vocabularies. The underlying similarity metrics

are the main object of research and are tuned to be well applicable within the DH community such that they deliver better results than the state of the art. A reference dataset is built and used for evaluation of the similarity algorithm and the validation and recommender system. A user evaluation is carried out to assess both systems as well as the user experience. In the first year of this work, the collection and evaluation of state of the art vocabulary editors and similarity algorithms has been conducted. In parallel, the basis of the enclosing vocabulary editor has been designed and implemented according to the results of a requirement analysis. The first version of the editor has been presented to users within and outside of the DH community. The comments and positive feedback of the participants strengthened the need for a simple to use, domain independent tool for building, curating and publishing vocabularies.

One challenge is to reach enough researchers for the recommender assessment and for giving valuable feedback. Another one is that all evaluation is carried out as neutral as possible to avoid tuning criteria to match a desired outcome. Among the countermeasures taken are increased cooperation with scholars and scientists, a continuous comparison of results with the state of the art and the use of reference datasets outside the DH domain.

Acknowledgments

This research is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG)—CRC 980 Episteme in Motion. Transfer of Knowledge from the Ancient World to the Early Modern Period. Project-ID 191249397, and supported by the Helmholtz Metadata Collaboration Platform and the German National Research Data Infrastructure (NFDI).

This preprint has not undergone peer review (when applicable) or any post-submission improvements or corrections. The Version of Record of this contribution is published in *The Semantic Web: ESWC 2022 Satellite Events*, and is available online at `https://doi.org/10.1007/978-3-031-11609-4_33`

References

1. Banerjee, S., Pedersen, T.: Extended Gloss Overlaps as a Measure of Semantic Relatedness. In: IJCAI-2003. pp. 805–810. Acapulco, Mexico (May 2003)
2. Chandrasekaran, D., Mago, V.: Evolution of Semantic Similarity - A Survey. ACM Computing Surveys **54**(2), 41:1–41:37 (Feb 2021). https://doi.org/10.1145/3440755
3. Gorman, J., Curran, J.R.: Scaling distributional similarity to large corpora. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. pp. 361–368. ACL-44, Association for Computational Linguistics, USA (Jul 2006). https://doi.org/10.3115/1220175.1220221
4. Han, M., Zhang, X., Yuan, X., Jiang, J., Yun, W., Gao, C.: A survey on the techniques, applications, and performance of short text semantic similarity. Concurrency and Computation: Practice and Experience **33**(5), e5971 (2021). https://doi.org/10.1002/cpe.5971

5. Haslhofer, B., Isaac, A., Simon, R.: Knowledge Graphs in the Libraries and Digital Humanities Domain. arXiv:1803.03198 [cs] pp. 1–8 (2018). https://doi.org/10.1007/978-3-319-63962-8_291-1
6. Hyvönen, E.: Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web **11**(1), 187–193 (Jan 2020). https://doi.org/10.3233/SW-190386
7. Jiang, Y., Zhang, X., Tang, Y., Nie, R.: Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. Information Processing & Management **51** (May 2015). https://doi.org/10.1016/j.ipm.2015.01.001
8. Lipp, J., Gleim, L., Cochez, M., Dimitriadis, I., Ali, H., Alvarez, D.H., Lange, C., Decker, S.: Towards Easy Vocabulary Drafts with Neologism 2.0. In: Verborgh, R., Dimou, A., Hogan, A., d'Amato, C., Tiddi, I., Bröring, A., Mayer, S., Ongenae, F., Tommasini, R., Alam, M. (eds.) The Semantic Web: ESWC 2021 Satellite Events. vol. 12739, pp. 21–26. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-80418-3_4
9. Nguyen, V., Yip, H.Y., Bodenreider, O.: Biomedical Vocabulary Alignment at Scale in the UMLS Metathesaurus. In: Proceedings of the Web Conference 2021. pp. 2672–2683. ACM, Ljubljana Slovenia (Apr 2021). https://doi.org/10.1145/3442381.3450128
10. Pastor-Sánchez, J., Martinez-Mendez, F.J., Rodríguez, J.: Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. Information Research, ISSN 1368-1613, Vol. 14, Nº. 4, 2009 **14** (Jan 2009)
11. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics **19**(1), 17–30 (Jan 1989). https://doi.org/10.1109/21.24528
12. Resnik, P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research **11**, 95–130 (Jul 1999). https://doi.org/10.1613/jair.514
13. Sánchez, D., Batet, M.: A semantic similarity method based on information content exploiting multiple ontologies. Expert Systems with Applications **40**, 1393–1399 (Mar 2013). https://doi.org/10.1016/j.eswa.2012.08.049
14. Sánchez, D., Batet, M., Isern, D.: Ontology-based information content computation. Knowledge-Based Systems **24**, 297–303 (Mar 2011). https://doi.org/10.1016/j.knosys.2010.10.001
15. Suominen, O., Mader, C.: Assessing and Improving the Quality of SKOS Vocabularies. Journal on Data Semantics **3**(1), 47–73 (Mar 2014). https://doi.org/10.1007/s13740-013-0026-0