Does it fit? KOS evaluation using the ICE-Map Visualization.

Kai Eckert[1], Dominique Ritze[1], and Magnus Pfeffer[2]

[1] University of Mannheim
University Library
Mannheim, Germany
{kai.eckert,dominique.ritze}@bib.uni-mannheim.de
[2] Stuttgart Media University
Stuttgart, Germany
pfeffer@hdm-stuttgart.de

Abstract. The ICE-Map Visualization was developed to graphically analyze the distribution of indexing results within a given Knowledge Organization System (KOS) hierarchy and allows the user to explore the document sets and the KOSs at the same time. In this paper, we demonstrate the use of the ICE-Map Visualization in combination with a simple automatic indexer to visualize the semantic overlap between a KOS and a set of documents.

1 Introduction

Hierarchical Knowledge Organization Systems (KOS), like thesauri, taxonomies, or other kinds of (lightweight) ontologies are widely used to describe all kinds of resources, large document corpora amongst others. In the Semantic Web, these KOSs are usually described in SKOS (Simple Knowledge Organization System[3]). The public availability of diverse KOSs on the web leads to new possibilities regarding the reuse of existing KOSs, but at the same time raises the question which KOS is suitable for the resources to be described. Thus, measuring the overlap of the subject coverage of a given document set and multiple KOSs is a necessary task before starting further, possibly time-consuming and costly efforts to annotate the documents. For these measurements, any one-dimensional analysis in the line of "summing the number of concepts that appear in the documents" is not sufficient. First, there is no baseline to compare the generated numbers with and second, all hierarchical information is lost and it is not possible to compare the results in their subject context. Instead, we propose to use a graphical visualization that preserves the hierarchical context as well as a statistical measure. The numerical results which are provided by the measure are intuitive to understand and well suited for a graphical representation. They are combined in the ICE-Map Visualization.

[3] http://www.w3.org/TR/skos-reference/

In this paper, we use the ICE-Map Visualization [2] to visualize the overlaps between KOSs and document sets. This visualization is based on a treemap and allows the user to browse a KOS hierarchy interactively. The colors indicate which parts of the KOS fit the documents. To create the visualization, the ICE-Map Visualization requires that the documents are annotated with KOS concepts. In the discussed use case, there are no annotations yet and manual assigment is obviously not feasible. Thus, it is necessary to automatically generate them. We show that the ICE-Map Visualization in combination with our automatic indexing approach is suitable to calculate and visualize the overlap between a KOS and a document set in a way that users can make informed decisions on whether the document set fits to the KOS.

2 Setup

We apply a KOS-based indexing approach to determine which concepts of the KOS occur in a given document. For this purpose, we developed a pure linguistic indexer called LOHAI[1] which is free and open source. It uses part-of-speech tagging, stemming, and word-sense disambiguation. It is especially important that the indexer does not rely on any additional knowledge sources and is kept simple to ensure usability as well as comprehensibility of the results. The reference implementation is available online[4]. The weighted concept annotations created by LOHAI form the basis for the ICE-Map Visualization.

The ICE-Map Visualization is an approach for visual datamining (VDM) specifically designed for the purpose of maintenance and use of concept hierarchies in various settings. In this paper, we use it to visualize the number of documents associated with the concepts in the thesaurus. The ICE-Map Visualization is described in detail by Eckert [2]. Here, we briefly recapitulate the basic idea and introduce the weight function employed in this paper.

The usage of a concept c is determined by a weight function $w(c) \in \mathbb{R}_0^+$ that assigns a non-negative, real weight to it. Based on this weight function, we further define:

$$w^+(c) = w(c) + \sum_{c' \in \text{Children}(c)} w^+(c') \tag{1}$$

with Children(c) being the direct child concepts (narrower concepts) of c. $w^+(c)$ is a monotonic function on the partial order of the concept hierarchy H, i.e., the value never increases while walking down the hierarchy. This gives the value of the root node root(c) a special role as the maximum value[5] of w^+, which we denote as $\hat{w}^+$: $\hat{w}^+(c) = w^+(\text{root}(c)) = \max_H w^+(c)$.

[4] `https://github.com/kaiec/LOHAI`

[5] The root node is defined as the only concept c in H for which holds that Parents(c) = $\emptyset$. Note that we require H to have a single root concept. Otherwise, we introduce an artificial single root concept that becomes the parent of all former root concepts.

If we use the number of annotations made for a given concept as the weight function $w(c)$, we can calculate the likelihood that a concept is assigned to a random document as follows[6]:

$$L(c) = \frac{w^+(c) + 1}{\hat{w}^+(c) + 1} \qquad L(c) \in (0, 1] \tag{2}$$

In information theory, the *information content* or *self-information* of an event x is defined as $-\log L(x)$, i.e., a higher information content means a more unlikely event. Together with a normalizing factor, we get the following definition for the information content $IC(c) \in [0, 1]$ of a concept c:

$$IC(c) = \frac{-\log L(c)}{\log(\hat{w}^+(c) + 1)} \qquad \hat{w}^+(c) \neq 0 \tag{3}$$

This is again a monotonic function on the partial order of H and assigns 0 to the root concept and 1 to concepts with $w(c) = 0$. The ICE-Map Visualization always visualizes the difference of two information contents based on two different weight functions or two different data sets: $D(c) = IC_1(c) - IC_2(c)$. The power of the ICE-Map Visualization lies in the possibility to choose arbitrary weight functions for IC_1 and IC_2. To calculate the weight of a concept regarding its usage in a document set, we use:

$$w_1(c) = \sum_{a \in \mathrm{Aset}(c)} \mathrm{Weight}(a) \tag{4}$$

with $\mathrm{Weight}(a)$ denoting the weight of a single annotation a as calculated by LOHAI[7] and $\mathrm{Aset}(c)$ being the set of annotations assigned to a concept c.

To evaluate the suitability, we compare the information content based on Equation 4 to the intrinsic information content [3] – a heuristic for the expected information content of a concept based on its position in the hierarchy. In our statistical framework, we obtain the intrinsic information content by employing the following weight function:

$$w_2(c) = |\mathrm{Children}(c)| \tag{5}$$

The ICE-Map Visualization uses a treemap to visualize the concept hierarchy together with the results of the analysis. It gives a broad overview of the whole document set with the annotated concepts and supports zooming and navigating the hierarchy of the KOS to get a detailed view. The automatic indexer LOHAI and the ICE-Map Visualization are included in our KOS analysis software SEMTINEL[8].

[6] The addition of 1 is necessary to allow a value of 0 for $w(c)$. Otherwise, the logarithm of $L(c)$ (cf. Equation 3) would not be defined for $w(c) = 0$.

[7] Strictly speaking, from an information-theoretic perspective, this function interprets the *tf-idf* weight of the annotation as the likeliness of being an annotation for the document. This interpretation is not correct, as *tf-idf* is no probability value.

[8] http://www.semtinel.org/

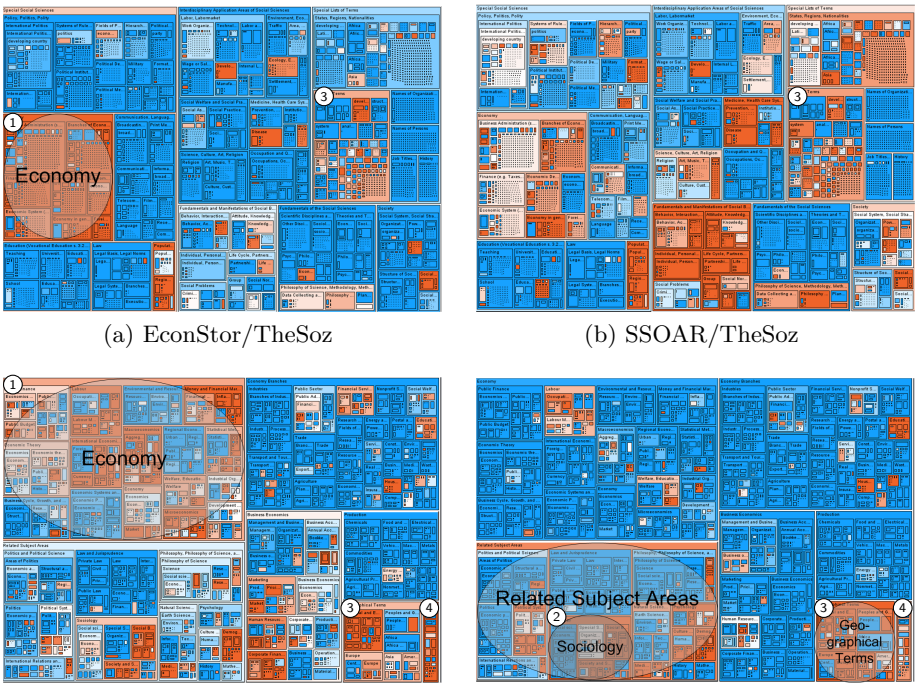
3 Experiments

To demonstrate the usefulness of the ICE-Map Visualization together with LOHAI to measure the suitability of thesaurus and document collection, comprehensive document sets and KOSs are needed. The KOSs need to have a significant overlap without describing the same topic and we also need at least one document set for each KOS where we can assume that it fits to the KOS. Furthermore, we would prefer to use well-established KOSs that are freely available and widely used. They need to have a significant size and at least one language in common.

For the experiments we chose TheSoz[9] (Thesaurus for the Social Sciences) version 0.86 and STW[10] (Standard Thesaurus Wirtschaft) version 8.08 in our experiments. Both KOSs are available as SKOS vocabularies and have a comparable size of about 7000 concepts with English labels. While TheSoz covers all social science disciplines, STW focuses on economical topics. As document sets, we apply SSOAR[11] and EconStor[12]. SSOAR as well as EconStor are open-access servers, maintained by GESIS and ZBW, respectively, the organisations that also publish the KOSs. Of both sets, we take a random subset of 2700 documents to ensure comparable results. As for the KOSs, SSOAR has its focus on social science and EconStor on economy. Despite of some deviations, we can assume that SSOAR naturally fits to TheSoz and EconStor fits to STW.

In Figure 1, we show the resulting visualization for all combinations of KOSs and document sets. The coloring represents the value of the weight function. It ranges from blue which means the weight for this concept is really low over white and finally to red which indicates a very high weight, compared to the reference weight determined by the heuristic. This economical bias of Econstor can clearly be seen in Figure 1a since most concepts which are used in the documents are narrower concepts of *Economy* ①. In contrast, the results of SSOAR/TheSoz (Figure 1b) do not point out such a clear focus on one specific field. It is interesting that Economy is still very visible, an indicator that both sciences indeed have an overlap reflected in the document sets. Moreover, the *General Terms* section ③ is used similarly by both document sets. When the STW is used as KOS, it can be seen in Figure 1c that EconStor documents contain concepts of several parts (especially *Economy* ①) while SSOAR documents use concepts which are narrower ones of *Related Subject Areas* and especially of *Sociology* (Figure 1d, ②). Other parts that are used well by both document sets are again general parts like *Geographical Terms* ③ and *General Terms* ④. All in all, the semantic overlaps of the document sets with the KOSs are clearly visible. Without any further information, we evaluated two document sets and two KOSs and were able to develop a deeper understanding of them by just browsing through the ICE-Map Visualization.

[9] http://lod.gesis.org/thesoz/
[10] http://zbw.eu/stw/versions/latest/download/about.en.html
[11] http://www.ssoar.info/
[12] http://www.econstor.eu/



(a) EconStor/TheSoz (b) SSOAR/TheSoz

(c) EconStor/STW (d) SSOAR/STW

Fig. 1: EconStor and SSOAR indexed with TheSoz and STW

4 Conclusion

We presented an approach to visualize the semantic overlap of a KOS and a document set. We combined the ICE-Map Visualization with a very simple automatic indexer called LOHAI. We chose two KOSs and two document sets with a significant topical overlap to demonstrate the usefulness of our approach. Based on the resulting visualization, we could show that it is possible to identify whether KOS and document set topically fit together. Thus, the choice of a suitable KOS or the maintenance of an already used KOS is strongly simplified.

References

1. Kai Eckert. LOHAI: Providing a baseline for KOS based automatic indexing. In *Proceedings of the first International Workshop on Semantic Digital Archives (SDA) at the International Conference on Theory and Practice of Digital Libraries (TPDL) 2011, Sep 29 2011, Berlin*, 2011.
2. Kai Eckert. The ICE-Map Visualization. Technical Report TR-2011-003, University of Mannheim, Department of Computer Science, 2011.
3. Nuno Seco, Tony Veale, and Jer Hayes. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In *Proceedings of the 16th European Conference on Artificial Intelligence*, pages 1089–1090. Valencia, Spain, 2004.