Estimating the relevance of search results in the Culture-Web: a study of semantic distance measures

Laura Hollink
Delft University of Technology
l.hollink@tudelft.nl

Mark van Assem
VU University Amsterdam
mark@cs.vu.nl

Abstract: More and more cultural heritage institutions publish their collections, vocabularies and metadata on the Web. The resulting Web of linked cultural data opens up exciting new possibilities for exploring these cultural heritage collections. We report on ongoing work in which we investigate the estimation of relevance in this Web of Culture. We study existing measures of semantic distance and how they apply to two use cases. The use cases relate to the structured, multilingual and multimodal nature of the Culture Web. We distinguish between measures using the Web, such as Google distance and PMI, and measures using the Linked Data Web, i.e. the semantic structure of and semantic links between vocabularies. We perform a small study in which we compare these semantic distance measures to human judgements of relevance. Although it is too early to draw any definitive conclusions, the study provides new insights into the applicability of semantic distance measures to the Web of Culture, and clear starting points for further research.

1 Introduction

More and more datasets are released on the Web, open for anyone to use and explore. This is happening in the government sector, the scientific sector, and also in the cultural heritage (CH) sector. Museums, archives and libraries are opening up not only their collections of artworks, but also the associated metadata and thesauri. Many CH institutions facilitate reuse of their data and linking to other collections by adhering to standards such as SKOS and Dublin Core. Several institutions have now published their data as Linked Open Data. The resulting Culture Web[1] opens up exciting new possibilities for searching and browsing through cultural heritage collections. A significant effort towards a Culture Web is made in the Multimedian E-Culture project[2], in which collections and vocabularies of several cultural heritage institutions are linked and made available on the Web.

In this paper we report on ongoing work in which we investigate the estimation of relevance in this Web of Culture. When is a search result relevant? The specific characteristics of the Web of Culture – its level of structure, multilingual and multimodal nature – ask for a re-evaluation of existing techniques. At the same time these characteristics enable new methods for estimating the relevance of search results. The cultural heritage sector has always had a large amount of structured background knowledge available. Collections

[1] Term first coined by Guus Schreiber in his inaugural speech, see `http://www.cs.vu.nl/~guus/talks/oratie/index.html`.

[2] Project: `http://e-culture.multimedian.nl/`. See `http://e-culture.multimedian.nl/resources/datacloud/` for an image of the subset of the Culture Web in this project.

are annotated by professional documentalists, using extensive but weakly structured vocabularies. Cultural heritage collections consist largely of items of a visual or physical nature: videos of television broadcasts, images of paintings, drawings, statues and cultural artifacts such as juwellery and china. Unlike the web of science, where English is the dominating language, the web of culture consists of many languages. Since the cultural objects are usually tied to the culture or history of a nation, annotation and search in the native language is natural and necessary.

One way to estimate relevance is with semantic distance measures, as for example applied by [Bro98, TH07, WSA+09]. Our aim is to contribute to the knowledge about how these measures perform on the Web of Culture. We study existing measures of semantic distance and how they apply to two use cases. The use cases relate to the structured, multilingual and multimodal nature of the Culture Web. We distinguish between measures using the Web, such as Google distance, and measures using the Linked Data Web, i.e. the semantic structure of and semantic links between vocabularies. To this end, we ask two human subjects to judge the relevance of search results to a query and compare these judgements to relevance estimates based on semantic distance. We will not be able to present any hard conclusions but our goal is rather to provide insights into the notion of relevance and to bootstrap future work in this area.

2 Related Work

Semantic distance measures can be classified into: (1) co-occurrence based measures which rely on a large corpus of text and (2) structural measures, which rely on a predefined semantic structure. Both types of measures show good results when compared to human assessments of semantic distance. It appears that the results can even be improved when the two types of measures are combined [MMR09]. Examples of the first type are Latent Semantic Analysis [LD97], Pointwise Mutual Information [Tur01] and Google Distance [CV07]. The latter is based on the ratio between the number of hits returned by Google for keywords x and y and the number of hits for a combined search for x AND y, normalized by the total amount of webpages indexed by Google. An example of the second type of measure is Leacock and Chodorow's [LC98] simLC, which measures similarity of concepts in a hierarchy as the shortest path between x and y, normalized by twice the maximum depth of the hierarchy. Other measures make use of, for example, the lowest common subsumer of two concepts or the information content of the concepts [BH01].

Several researchers have exploited the availability of linked data on the web, and the application of different variants of structural semantic distance measures to retrieval or recommendation of relevant items is getting renewed attention. SemRank [AMS05] uses a mix of semantics to rank results for conventional or discovery oriented searches. Ruotsalo et al. [TH07] combine the semantic distance between concepts over several relations. Wang et al. [WSA+09] analyze which relationships between artworks deliver the best recommendations. The task we study differs from Ruotsalo and Wang in that we target search instead of recommendation, i.e. we start with a query instead of an annotated (and possibly rated) work of art. We set out to explore how semantic distance measures can be applied to estimate relevance in the Culture Web. As a first step, we implemented two established measures and tested them in two use cases: Leacock and Chodorow [LC98] as an

example of a structural measure and Normalized Google Distance (NGD) as a Web-based co-occurrence measure.

3 A comparative study

We have selected two highly specific but typical situations to study the applicability of semantic distance measures in the Web of Culture. Semantic distance is generally measured between concepts or words. However, in the Web of Culture a user is often interested in the relevance of an image that is annotated with these concepts or words. This potentially alters the way semantic distance measures can be relied on. We have tested this situation using Dutch and English language queries and annotations, taken from domain specific and general purpose semantic resources. Both use cases are performed with vocabularies and collections of images of art objects from the Multimedian E-culture project.

Use Case 1: Ethnological artifacts and SVCN The first use case is search in the collection of artifacts from the Dutch National Museum of Ethnology. The artifacts are annotated with a facetted thesaurus, the SVCN. The annotations are of high quality, made by experts. They include the culture, region and period from which the artifact originates, and the type of object it concerns. For this use case, 21 queries were drawn randomly from SVCN's Object facet, such as "tools for spinning and plying", "flutes", "membranophones: struck". These examples show that SVCN concepts can be highly domain specific.

Use Case 2: Rijksmusem, AAT and WordNet In the second use case we study the situation in which queries are from a different vocabulary than the annotations. We search a collection of images of artifacts in the Dutch Rijksmuseum, which is annotated with object types from the Art and Architecture Thesaurus (AAT). Each AAT concept has one or more English labels. The annotations are of lower quality; the artifacts were originally annotated with an in-house vocabulary and later automatically translated to the AAT in the Multimedian E-Culture project [SAvA+06]. We selected queries from WordNet. It's widespread use and general purpose concepts makes it an interesting test case. WordNet is a freely available thesaurus of the English language. Each concept has multiple synonymous labels. We drew 24 query concepts from the Physical object branch of the hierarchy, with labels such as "pot", "feather boa", "ring armor" and "artillery".

Evaluation of two semantic distance measures We implemented two established semantic distance measures and tested them on the two use cases: simLC by Leacock and Chodorow as an example of a structural measure and Normalized Google Distance as a Web-based co-occurrence measure. SimLC actually measures semantic proximity of concepts: the higher the score, the more similar two concepts are. NGD measures the distance between words or, in our case, labels. We slightly adapted NGD in the sense that we limited our Google searches to Dutch language pages in the first use case, and English language pages in the second.

To evaluate the two measures, we created a test set. Two human raters were presented with the same series of queries and resulting images of artifacts. They were asked to evaluate the

relevance of the images according to the following question: "How relevant is this work of art as a search result for the given query?". Ratings were on a 6-point scale ranging from "highly irrelevant" to "highly relevant". Rating was done with a web-based application displaying (1) a query including a definition and synonymous terms if available and (2) a search result in the form of an image of an artifact with a description. The two raters were not familiar with the art domain, but were habitual users of the Web. Although the raters sometimes disagreed on the judgements, all conclusions presented below are valid for both raters (with one small exception, see section 4). We tested inter-rater agreement with a weighted Cohen's κ [Coh68], which was reasonably high: 0.70 (squared weights).

We determined the correlation between the computed semantic distance and the human judgements with Spearman's ρ [Spe04]. This measure ranges from -1 to 1; values close to 0 mean a poor correlation and thus a poor estimation of relevance, values closer to -1 or +1 denote a stronger correlation and thus a good estimation of relevance.

The human raters judged relevance of query-artifact pairs. However, one artifact can be annotated with multiple concepts. As a result, multiple semantic distance scores between a query-concept and an annotation-concept need to be combined to come to an estimation of the relevance of an artifact for a query. This effect is even stronger for NGD since this measures the distance between words (or labels), not concepts. Query-concepts and annotation-concepts can both have multiple labels, resulting in a number of NGD measures for each query-artifact pair. We studied three methods to combine distance scores: (1) the mean of the scores, (2) the maximum of the scores, and (3) selecting only the score(s) of the annotation-concept that denotes the type of object depicted in the image, leaving out other annotations such as creator, location or culture. This part of the annotation comes closest to the queries, since all queries are for object types (e.g. "teapot", "firearm").

4 Results

Use Case 1: Ethnological artifacts and SVCN Table 1(a) shows the correlation between the human relevance judgements and estimations of relevance based on simLC scores for query–annotation pairs. The different ways to combine the semantic distance scores have a clear impact on the quality of the estimations. Taking the maximal score works better than taking the mean of the scores. A possible explanation is that each artifact is annotated with a few concepts that have no similarity at all with the query, such as the creator or the country of origin. Taking the mean score unjustly lowers the estimated relevance. The only method that leads to reasonable correlation is to select only object type annotations. This shows that it is beneficial to take into account what kind of query is posed (a person, location, object type, etc.) and to use knowledge of the metadata schema to pre-select suitable parts of the annotation.

We applied Normalized Google Distance to the same use case; results are shown in table 1(b). Contrary to simLC, the mean of the scores leads to a better estimate of relevance than the maximum of the scores. We assume this is because NGD measures co-occurrence rather than similarity. Using only the object type annotations again brings the best performance. Overall the relevance estimations of NGD are lower than those of simLC and the observed correlations are not statistically significant. Our hypothesis is that this is due to the domain specific, rare nature of our concepts and their labels, which is aggravated by

the fact that they are in Dutch. This leas to a low number of hits in Google and unreliable distance scores. NGD estimates are reasonable for relatively common query and annotation terms such as svcn:footgear by form — svcn:sandals and svcn:kit bag — svcn:sling (as in baby sling). They are completely off for terms like holders for chewing tobacco. We checked if the low performance was caused by labels consisting of multiple words, but we did not find such an effect.

Use Case 2: Rijksmusem, AAT and WordNet Table 1(c) show the results of use case 2. We see a clear rise in the performance of NGD as compared to use case 1. A plausible reason is that this use case contains queries in English from the general purpose resource WordNet. In line with this hypothesis we again observe that NGD works better for relatively common query words (e.g. "pot") than for rare words (e.g. "caldron") or ambiguous words (e..g "shovel", meaning a type of artillery). In the Rijksmuseum collection that was used for this use case, the only structured metadata that we had access to were object type annotations. Each annotation concept and each query had multiple labels. Table 1(c) therefore shows two different ways to combine the NGD scores for the labels of the object type annotations: mean and max. Again the mean of all NGD scores gives a better estimate of relevance than the maximum of all NGD scores of a query-result pair. As this use case consists of two separate vocabularies, standard structural similarity measures are not applicable.

(a)

Use case 1	**simLC**
mean	0.19*
max	0.25*
object type	0.45*

(b)

Use case 1	**NGD**
mean	-0.08
max	-0.02
object type	-0.19

(c)

Use case 2	**NGD**
object type mean	-0.25*
object type max	-0.16†

Table 1: Correlation between human relevance judgements and estimations of relevance the measures simLC and NGD. A * denotes significance at the 0.05 α-level; † denotes a significant correlation with the judgements of only one of the two raters.

5 Discussion and future work

In this paper we have applied two basic measures of semantic distance to two use cases in the Web of Culture. This exercise has provided us with some valuable insights into the relation between semantic distance and the notion of relevance on the Web of Culture. The applicability of purely web-based measures such as Normalized Google Distance seems to be limited in this context. The terms and their interpretation are so domain specific that word sense disambiguation errors are common. The rareness of the terms and the fact that the Web of Culture consists of many smaller languages leads to very low numbers of hits for google queries, which in turn leads to unreliable distance scores. One interesting direction for future research is to try other, more domain specific corpora for word co-occurrence based measures. The general Web might not be the best choice in this case.

Structural measures seem to work, but only when implemented with knowledge of the collection, metadata and queries. This means that semantic distance measures cannot be used

off the shelve. We plan future research into how to incorporate this type of knowledge in semantic distance measures, and moreover, how to (semi) automatically accumulate this knowledge. There is currently only little evidence that structural semantic distance measures work also across vocabularies. With the growth of open linked data this is likely to become more and more demand for measures that work across vocabularies and datasets.

These insights into the behavior of semantic distance measures on the Culture Web can be used to take the next step and design meaningful measures that combine the strengths of measures based on the Web or a domain specific corpus with structural measures, perhaps taking into account knowledge about the collection.

References

[AMS05] K. Anyanwu, A. Maduko, and A. Sheth. SemRank: Ranking Complex Relationship Search Results on the Semantic Web. In *Proc. of the 14th Intl. Conf. on World Wide Web*. ACM Press, 2005.

[BH01] A. Budanitsky and G. Hirst. Semantic distance in WordNet: an experimental application oriented evaluation of five measures. In *Proc. of the NACCL Workshop: on WordNet and other lexical resources: Applications, extensions, and customizations*, pages 29–34, 2001.

[Bro98] T. A. Brooks. The Semantic Distance Model of Relevance Assessment. In *Proc. of the 61st Annual Meeting of ASIS*, pages 33–44, 1998.

[Coh68] J. Cohen. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 20:213–220, 1968.

[CV07] R. Cilibrasi and P. Vitanyi. The google similarity distance. *IEEE Transactions on knowledge and data*, 19(3):370–383, 2007.

[LC98] C. Leacock and M. Chodorow. *WordNet:An Electronic Lexical Database*, chapter Combining Local Context and WordNet Similarity for Word Sense Identification, pages 265 – 285. MIT Press, 1998.

[LD97] T. Landauer and S. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psych. Review*, 104(2):211–240, 1997.

[MMR09] Y. Marton, S. Mohammad, and P. Resnik. Estimating Semantic Distance Using Soft Semantic Constraints in Knowledge-Source-Corpus Hybrid Models. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pages 775–783, 2009.

[SAvA+06] G. Schreiber, A. Amin, M. van Assem, V. de Boer, L. Hardman, M. Hildebrand, L. Hollink, Z. Huang, J. van Kersen, M. de Niet, B. Omelayenko, J. van Ossenbruggen, R. Siebes, J. Taekema, J. Wielemaker, and B. Wielinga. MultimediaN E-Culture demonstrator. In *Proc. of the 5th Int'l Semantic Web Conf.*, number 4273 in Lect. Notes in Computer Science, pages 951–958, 2006.

[Spe04] C. Spearman. The proof and measurement of association between two things. *Amer. J. Psychol*, 15:72–101, 1904.

[TH07] T.Ruotsalo and E. Hyvönen. A method for determining ontology-based semantic relevance. In *Proc. of the Intl. Conf. on Database and Expert Systems Applications*, pages 680–688, 2007.

[Tur01] P. Turney. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proc. of the 12th European Conference on Machine Learning*, pages 491–502. Springer, 2001.

[WSA+09] Y. Wang, N. Stash, L. Aroyo, L.Hollink, and G. Schreiber. Using Semantic Relations for Content-based Recommender Systems in CH. In *Proc. of the Workshop on Ontology Patterns at ISWC*, 2009.