

Usage Contexts for Object Similarity: Exploratory Investigations

Katja Niemann, Hans-Christian Schmitz,
Maren Scheffel, Martin Wolpers
Fraunhofer Institute for Applied Information Technology (FIT)
Schloss Birlinghoven
Sankt Augustin, Germany
{katja.niemann, hans-christian.schmitz,
maren.scheffel, martin.wolpers}@fit.fraunhofer.de

ABSTRACT

We present new ways of detecting semantic relations between learning resources, e.g. for recommendations, by only taking their usage but not their content into account. We take concepts used in linguistic lexicology and transfer them from their original field of application, i.e. sequences of words, to the analysis of sequences of resources extracted from user activities. In this paper we describe three initial experiments, their evaluation and further work.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Content Analysis and Indexing – *Linguistic processing*, Information Search and Retrieval – *Clustering*.

General Terms

Algorithms, Measurement, Experimentation, Verification

Keywords

Text Mining, Clustering, Usage Data Analysis, Usage Contexts

1. INTRODUCTION

Analyzing the behavior of learners and deducing something meaningful from it is a major obstacle for teachers. In recent times, more and more educational institutions have therefore started to use electronic means to manage learning as well as teaching activities. A tool frequently used for this purpose is a Learning Management System (LMS). LMSs can handle activities from learners and teachers as well as administrative processes, ranging from learning resource provision to course related management tasks and grading. Many recent LMSs also include communication tools, such as chat, group discussion and email. In many cases a LMS collects data about the users' interaction with information and services by the system, thereby generating large amounts of data. So far, these data are not made use of to their full extend, e.g. individual analysis of a student's learning behavior or performance. Both analysis results, however, could be used to improve the learning efficiency and effectiveness. For example, a teacher being notified about a student's continuous search activities could address that student and help her personally. Furthermore, the system might let the teacher know that many students within one course follow the same search pattern which could be an indication for the teacher to provide additional learning resources. An approach to extract representative activity patterns from usage data is described in [1].

With the learner's activities identified, recommendations of suitable resources, i.e. learning resources that are thematically related to a learner's current usage context. are often difficult to apply or even fail: while the manual creation of semantic metadata is costly, automatic extraction of semantic metadata only works well for text-based resources. LMSs however often also have to deal with media types such as images, audio and video files. Adding social metadata to resources is another way to classify objects based on their content but it tends to be ambiguous or even faulty.

In this paper, we present new ways of extracting semantic relations between learning resources by analyzing data of the users' interaction with the system and thus the usage of learning resources. In contrast to collaborative filtering approaches [2] our approaches do not rely on the relations between users and objects but on the relations between the objects themselves, i.e. they are used together or used in similar contexts. We describe our first exploratory analyses of applying semantic relation detection techniques from corpus-driven lexicology to semantic object relation detection in learning contexts. The datasets of usage activities collected within the MACE portal [3] (a thematic web-based portal for learning resources in architecture) serve as our test bed.

The rest of the paper is structured as follows: In section 2 we report on the background of similarity calculation in recommender systems. In section 3 we describe three ideas of how to adopt lexicological techniques to identify related learning resources. We then present the corresponding experimental investigations using the MACE-LOR [3] as a test-bed in order to illustrate our ideas in section 4 and give a summary and an outlook in section 5.

2. BACKGROUND

Recommender systems deal with the delivery of items selected from a collection that the user is likely to find interesting or useful. The most common approaches of object similarity calculation used in recommender systems are user- or item-based collaborative filtering, content-based filtering or hybrid approaches which combine aspects of more than one filtering technique [4]. Commonly, hybrid systems combine content-based and collaborative-based techniques, but knowledge-based or demographic-based techniques can for example be integrated as well [5, 6].

Recommender systems in TEL need to satisfy other requirements than for example recommenders in e-commerce. The recommended items need to fit into the current context of the user, considering e.g. her competences and learning goals to ideally support her while learning. Approaches that deal with those requirements are comprehensively described in [7]. Ruiz-

Iniesta et al. [8] introduce a domain ontology holding interrelated concepts to index learning objects and establish learning paths among them. The approach of Bozo et al. [9] relies on teacher profiles, e.g. information about educational level, subject, area, region and school type, and on learning object profiles, e.g. information about educational level and topic. Those systems are well adapted to the domain of technology enhanced learning. However, they have in common that they need explicit information about the items and/or the users, which is, in practice, usually sparse or not available at all.

We adapt the item-based collaborative filtering approach but use a different way of calculating item similarities. For two objects to be deemed similar, their context of usage needs to be similar. We presuppose that users in TEL scenarios work predominantly on only one task within one session. That is, if two objects are used within the same context then they are most probably related via the session task, and this relatedness is a semantic/content-induced one.

3. WORD CONTEXTS AND SEMANTIC RELATEDNESS

If you want to know the meaning of a word, it is a good strategy to look it up in a dictionary where you might find a helpful definition. In many cases, however, a definition (if there is one at all) might not be sufficient to fully understand and thus correctly use the word. The words “strong” and “powerful”, for example, have highly related meanings. Yet, we can say “strong tea” while we cannot say “powerful tea”. “Powerful drug” though is acceptable [10]. Definitions of the word’s meaning will most probably not cover such differences. Therefore, dictionaries usually give contexts in which a word typically occurs to illustrate the actual word usage.

Context is considered to be significant for the meaning of a word : “You shall know a word by the company it keeps.” [11]. The company a words keeps – its co-occurring words – contributes to the meaning. Two words might just co-occur by accident. However, the co-occurrence might also be relatively frequent and thus statistically significant. Statistically significant co-occurrences reveal close relationships between the co-occurring words or their meanings respectively: they are used to detect multi-word expressions (‘New York’), idioms (‘kick the bucket’ [12]) or constructions with a milder idiomatic character (‘international best practice’ [10]).

Subsequently the question arises whether we can apply that insight to learning resources and their usage contexts: do significant co-occurrences of learning resources in learning contexts reveal close semantic relationships between the resources? If two words occur in very different contexts, they can be assumed to be semantically non-related. If one word occurs in various contexts, i.e. contexts of different types, we can assume the word to be polysemous that is different contexts correspond to different readings of the word. If two words, however, occur in very similar or even identical contexts, then we can assume that these words are semantically strongly related. ‘Relatedness’ can be specified as a kind of similarity: if, for example, the two words are co-hyponyms (that is, they have a common superordinate concept, which is very often true for words with highly similar contexts), they are similar regarding their superordinate concept [13, 14]. (Note that context-similarity is different from co-occurrence. Words with similar contexts do not need to co-occur in the same contexts.)

Thus, by comparing the usage contexts of words we can detect semantic similarity. It is now very interesting to find out whether this is true also for objects other than words: can we detect semantic similarities between learning resources by comparing the usage contexts of these objects? This seems to be intuitively plausible but has to be proven.

Words co-occur with other words, with some of them a statistically significant number of times. The significant co-occurrences form the co-occurrence class of the respective word. It can now be examined whether words significantly co-occur in cooccurrence classes. These words again form another co-occurrence class, namely a higher order co-occurrence class. After several iterations higher order co-occurrence classes become semantically homogenous. Heyer et al. [14] show this for the cooccurrences of ‘IBM’, among other words. Their investigations are based on text

corpora collected for the portal wortschatz.uni-leipzig.de (concerning the German treasury of words). The first co-occurrence class is rather heterogeneous, containing words like ‘computer manufacturer’, ‘stock exchange’, ‘global’ and so on. After some iterations of computing higher order co-occurrences classes, however, the classes become more homogenous and stable. The co-occurrence class of tenth order only contains names of other computer-related companies like ‘Microsoft’, ‘Sony’ etc. We can easily compute higher order co-occurrence classes for learning resources by deriving the first co-occurrences from usage contexts. The question is: do these classes become semantically homogenous, like the analogue classes of words?

Let us take stock: current lexicology heavily relies on the investigation of word contexts. By comparing entire contexts and examining co-occurrences and higher order co-occurrences semantic relations can be detected. We adopt the notions from lexicology and test whether they can be fruitfully applied in information retrieval for the analysis of learning activities.

4. EXPERIMENTS

We performed three initial experiments for answering the questions asked above:

- Can we detect semantic similarities of learning resources by comparing the usage contexts of these resources?
- Do significant co-occurrences of data objects in usage contexts reveal close semantic relationships between the learning resources?
- Do higher order co-occurrence classes of data objects become semantically homogenous, like the analogue classes of words?

We tentatively answer the three questions with ‘yes’ and take our answers to be hypotheses that are to be validated.

The MACE-portal serves as a test bed for our experiments. MACE (*Metadata for Architectural Contents in Europe*, [3]) relates architectural learning resources stored in various repositories with each other to support students in finding relevant information. To enable connections between these learning resources, metadata representations of the resources are stored in the central MACE repository. The representations base on the MACE application profile which extends the Learning Object Metadata (LOM) standard [15] for architecture-specific needs.

Our approach only uses the metadata representations of the learning resources: The attributes stored in the LOM instances are used to build document vectors for all learning resources. To this end, the English titles and descriptions of the learning resources,

their repositories, their learning resource types as well as their free text tags, classifications and assigned competences are taken into account. The semantic similarity between two learning resources is then calculated using the cosine similarity measure [16] between their respective document vectors. The cosine similarity measure forms the baseline for evaluating similarities derived from the analysis of usage contexts.

Within MACE, usage data are collected and stored using the CAM (Contextualized Attention Metadata [17, 12]) schema. A CAM instance comprises amongst others the user, the accessed resource, and the action performed on that resource. All CAM instances can be assigned to user sessions. For our experiments, we assume that a session constitutes the usage context for data resources and comprises all resources accessed by a user without a break longer than an hour. Altogether we consider 3130 sessions in which 11429 resources were accessed.¹

4.1 Experiment 1: Context Similarities

Our first hypothesis is that learning resources with similar usage have similar content. Before we can test the hypothesis, we have to define ‘usage’ and ‘similar usage’. We do so via the concept of a Usage Context Profile (UCP): two resources have similar usage iff they have similar UCPs. The UCP of a resource is the set of its usage contexts. Usage contexts are related to user sessions: we define a usage context of a resource R as a pair $\langle pre, post \rangle$, where pre is the bag of resources that a particular user accessed before R in the same session, and $post$ is the bag of resources that the user accessed after R in the same session. Pre- and post-contexts can be represented as vectors, with the resources as dimensions and the numbers of accesses as coordinates. Fig. 1 shows the UCP of a resource E . E occurred in two sessions: in the first session it was accessed after A (pre-context) and before D and C (postcontext); in the second session it was accessed after C and F and before C and G . That is E ’s UCP, consisting of two usage contexts which in turn both consist of a pre- and a post-context.

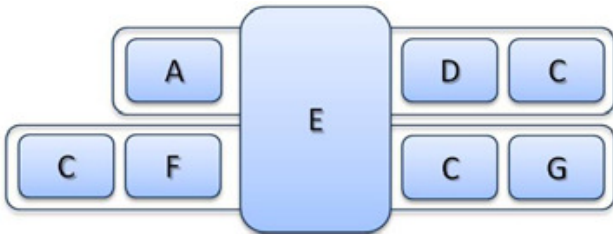


Figure 1. Usage Context Profile for resource E

Next we define a measure of similarity for UCPs. The similarity of two UCPs is defined as the arithmetic mean of the pair-wise similarities of their respective usage contexts. The similarity of two usage contexts is defined as the arithmetic mean of the similarities of the associated pre- and post-contexts. Finally, the similarity of two preor post-contexts is defined as the cosine similarity of the vectors representing these contexts. Note that usage similarity as it is defined here does not require conjoint usage – two learning resources are similar regarding their usages iff they have similar usage contexts. This does not require that they are ever used in the same session. So far, we do not take the order of learning resources in pre- and post contexts into account.

By the concept of UCP-similarity we define a means for calculating usage similarity values for resource pairs. With respect to the two kinds of similarity values we can now refine our initial hypothesis and make it testable. The new hypothesis is: resources with similar usages have similar content, therefore the usage similarity values of resource pairs are positively correlated with their semantic similarity values.

To test the hypothesis we generated two arrays of similarity values, one for the usage similarities of MACE learning resources and one for the semantic similarities of these resources. Thereafter, we computed the Pearson Correlation Coefficient [19] for these arrays. The result was a correlation coefficient of 0.35. That is, we detected a medium positive correlation between the different kinds of similarities. As our sample of resources was sufficiently large, the correlation coefficient can be regarded as representative. However, we expected a strong positive correlation and therefore our hypothesis is not strongly supported.

We know that our semantic metadata from which the semantic similarity values were computed suffer from a scarcity problem. Most probably, this affected the result of the evaluation. We therefore manually compared the 100 resource pairs with the highest usage context similarity values. 92% of these resource pairs showed content similarities. We found text documents on the same topic, e.g. “risk factor analysis”, or resource pairs with one of them being an exercise and the other one a text on the same topic, among other kinds of similarities. We also discovered that many of the detected content similarities are not entailed in the semantic metadata and are thus not accounted for in our evaluation. We therefore expect that a better baseline derived from a richer set of semantic metadata will lead to a stronger correlation of usage and content similarities. For a detailed description of the approach and its evaluation see [20].

4.2 Experiment 2: Conjoint Usage, First Order Co-occurrence

Our second hypothesis is that resources that are significantly often used together are semantically related. Semantically related objects can for example be about the same topic, they can be complementary regarding learning goals, etc. Since we access the semantics of learning resources via their metadata, we presume that all forms of semantic relatedness are somehow reflected as similarities of the metadata.

We test our hypothesis as follows: We relate resources regarding their conjoint usage in sessions, thereby taking not only the number of conjoint sessions but also the number of accesses within these sessions into account. We normalize the values of conjoint usage by referring to relative usage frequencies. Based on these values we construct a graph with resources as nodes and normalize conjoint usages as weights of edges. We then apply a graph clustering algorithm. A common approach for graph-based clustering is Markov Clustering (MCL, [21]). However, since there is no guarantee that the MCL algorithm terminates, we apply Iterative Conductance Cutting (ICC) for clustering the graph which is similar to MCL regarding scalability [22]. The ICC algorithm starts with only one cluster and iteratively splits a cluster into two new clusters until the performance measure (conductance) is below a specific threshold. The conductance represents the relation of cross-cluster edges to cluster internal edges. In other words, this approach gives greater importance to vertices which have many similar neighbours and less importance to vertices which have few similar neighbors. The iteration ends

¹ If you are interested in the data set, please contact the authors

when no more clusters can be split without violating the threshold.

Since we claim that conjoint usage gives rise to semantic relatedness and that semantic relatedness is reflected in the similarity of semantic metadata, we expect that the resulting clusters are semantically dense. In other words: we expect the members of each cluster to be semantically similar, while members of different clusters are semantically distinct. We compute semantic similarity as metadata similarity; the semantic density of a cluster is the average semantic similarity of its members. We refine our initial hypothesis as follows: the mean semantic similarities (metadata similarities) of the individual clusters systematically differ from the mean semantic similarity of the whole population. We expect the mean similarities of the clusters to be significantly higher than the corresponding mean similarity of the entire set of objects.

For the mean similarity value of each cluster we tested whether it significantly deviates from the mean of the entire population by applying a t-test. The results showed that 78% of the 106 generated conjoint-usage clusters have a significantly over-average semantic relatedness ($p < 0.05$) compared to the entire population. Only about 7% of the clusters had a significantly under-average mean similarity.

This result is very promising. It clearly supports the hypothesis that conjoint usage gives rise to semantic relatedness.

4.3 Experiment 3: Higher Order Co-occurrence

Our third hypothesis is that usage-based higher order co-occurrence classes of learning resources are semantically homogenous. We presume that semantic homogeneity correlates with semantic similarity. That is, we expect that the members of a semantically homogenous class are more similar than the members of a class of randomly chosen resources.

The concept of higher order co-occurrences is illustrated in Fig. 2. The objects *A*, *B* and *E* directly co-occur with object *C* in at least one session, while the objects *B*, *D* and *F* directly co-occur with object *A* and the objects *D* and *G* directly co-occur with object *E*. Therefore, objects *B*, *D*, *F* and *G* are second order co-occurrences of object *C* and form a second order co-occurrence class.

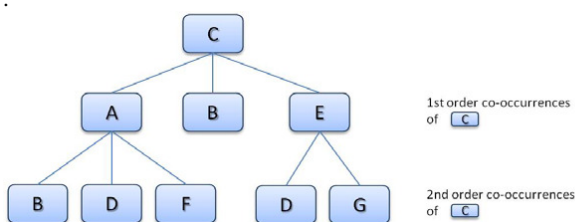


Figure 2. Co-occurrence tree of resource *C*

We recursively compute significant co-occurrences of resources and thus generate higher order co-occurrence classes of resources. We do so by taking the sessions as input to generate significant (first order) co-occurrences and use these co-occurrence classes as input for the calculation of the second order co-occurrence classes which then form the input for the calculation of the significant third order co-occurrence classes and so forth.

For simplicity reasons, we consider all co-occurrences of a resource in the example in Fig. 2. In practice, we are only interested in significant co-occurrences. Therefore, we need to

calculate a significance value for each resource in each cluster and to define a threshold. Resources with a significance value lower than the threshold are deleted from the classes and are not considered in the next calculation step. The significance value of a resource *O* in a co-occurrence class for a reference resource *R* is calculated by relating the number of contexts (here: sessions) containing *O* and *R* respectively, the number of contexts holding both resources and the total number of contexts. For further details see [14]. Currently, there is no generally accepted approach to define the significance threshold. Thus we choose an exploratory procedure for this problem as well and experiment with different thresholds.

The computation of the co-occurrence classes stops when the classes stabilize. In our experiment this happened after six iterations. Given that this approach is not a hard clustering an object can be contained in more than one class.

We evaluated the semantic density of the generated clusters as we did in experiment 2 using a t-test to compare the mean similarity of each cluster with the mean similarity of the population. Thereby, we gathered a significant over-average similarity of 82% of the 184 generated clusters ($p < 0.05$) compared to the mean similarity of all resources.

In a manual inspection of the higher order co-occurrence classes we found some interesting relations between the resources, e.g. a cluster containing only Spanish documents describing building material, mainly cement, concrete and bricks; or another cluster containing only English documents describing modern, public buildings, mainly museums and police stations.

5. OUTLOOK

We presented approaches for the identification of related learning resources from usage behavior exploring techniques borrowed from lexicology to relate learning resources by only focusing on their usage contexts. If these techniques are successful, they can be fruitfully applied for recommender systems. We started with exploratory, experimental investigations. The results clearly indicate that the chosen approach is promising.

We will investigate the appropriate parameters and thresholds that are used by the presented approaches. To this end, we will conduct additional experiments in other test beds in order to explore domain-dependent distinctions that need to be addressed. Moreover, we will adopt further natural language processing and information retrieval technologies and apply them to the objects' usage contexts. The usage contexts of an object *O* can be subsumed under one vector holding all objects the object *O* was ever accessed with. Such representations can be used as input for item-based collaborative filtering approaches, where an object is normally described by a vector holding all persons that, for instance, viewed or rated the object. Assuming objects can be handled as words, we can also conduct techniques like Latent Semantic Analysis for dimension reduction on these vectors before using it for similarity calculations.

Furthermore, we will explore if the presented approach can be transferred from relations between documents to relations between humans. For example, pre- and post contexts can be based on one's own activities or that of others, etc. This way, we would be able to identify relations among humans that are expressed and quantified by their behavior.

In order to test the approaches in real world settings, we will embed the techniques in recommender systems in the area of TEL. For example, within the MACE portal, a recommender could provide resources thematically related to the ones used in the

current session. Evaluations of this recommender will show if the chosen approach yields acceptable results.

6. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no 231396 (ROLE project).

7. REFERENCES

- [1] Scheffel, M., Friedrich, M., Niemann, K., Kirschenmann, U., Wolpers, M.: A Framework for the Domain-Independent Collection of Attention Metadata. In: Wolpers, M. et al. (eds.): Sustaining TEL: From Innovation to Learning and Practice. 5th European Conference on Technology Enhanced Learning, EC-TEL 2010. Proceedings, Lecture Notes in Computer Science, Volume 6383, pp. 426-431 (2010)
- [2] Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering* 17 (6), pp.734-749 (2005)
- [3] Stefaner, M., Dalla Vecchia, E., Condotta, M., Wolpers, M., Specht, M., Apelt, S., Duval, E.: MACE - enriching architectural learning objects for experience multiplication. In: Duval, E., Klamma, R., Wolpers, M. (eds.): *Creating New Learning Experiences on a Global Scale*, LNCS, vol. 4753, Springer, Heidelberg, pp. 322-336 (2007)
- [4] Candillier, L., Jack, K., Fessant, F., Meyer, F.: State-of-the-Art Recommender Systems. In: Chevalier, M. et al. (eds.): *Collaborative and Social Information Retrieval and Access – Techniques for Improved User Modeling*. Idea Group Publishing, pp. 1-22 (2009)
- [5] Burke, R.: Hybrid recommender systems: Survey and experiments. In: *The Adaptive Web* (2007).
- [6] Melville, P., Mooney, R.J., and Nagarajan, R.: "Content-Boosted Collaborative Filtering for Improved Recommendations". In: *Proc. of the 18th National Conference for Artificial Intelligence* (2002).
- [7] Manouselis, N., Vuorikari, R., van Assche, F.: Collaborative recommendation of e-learning resources: an experimental investigation. In: *Journal of Computer Assisted Learning*, 26, pp. 227-242 (2010)
- [8] Ruiz-Iniesta, A., Jimnez-Daz, G., Gmez-Albarrn, M.: User-Adaptive Recommendation Techniques in Repositories of Learning Objects: Combining Long-Term and Short-Term Learning Goals. In: *Proceedings of EC-TEL 2009*, LNCS 5794, pp. 645-650 (2009)
- [9] Bozo, J., Alarcón, R., Iribarra, S.: Recommending Learning Objects According to a Teachers Context Model. In: *Proceedings of EC-TEL 2010*, LNCS 6383, pp. 470-475 (2010)
- [10] Manning, C., Schütze, H.: *Foundations of Statistical Natural Language Processing*. MIT Press (1999).
- [11] Firth, J.: *A Synopsis of Linguistic Theory 1930–55*. *Studies in Linguistic Analysis*, Oxford: The Philological Society (1957).
- [12] Evert, S.: Corpora and collocations. In: Lüdeling, A., Kytö, M. (eds.): *Corpus Linguistics. An International Handbook. Volume 2*, de Gruyter, Berlin (2009)
- [13] Hoey, M.: Corpus linguistics and word meaning. In: Lüdeling, A., Kytö, M. (eds.): *Corpus Linguistics. An International Handbook. Volume 2*, de Gruyter, Berlin (2009)
- [14] Heyer, G., Quasthof, U., Wittig, T.: *Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse*. W3L-Verlag, Herdecke (2006)
- [15] LOM IEEE Standard for Learning Object Metadata, IEEE Std 1484.12.1 (2002)
- [16] Tan, P.-N., Steinbach, M., Kumar, V.: *Introduction to Data Mining*. Addison-Wesley, ISBN 0-321-32136-7, pp. 500 (2005)
- [17] Wolpers, M., Najjar, J., Verbert, K., Duval, E.: Tracking Actual Usage: the Attention Metadata Approach. *Educational Technology & Society* 10(3), pp. 106-121 (2007)
- [18] Schmitz, H.-C., Kirschenmann, U., Niemann, K., Wolpers, M.: Contextualized Attention Metadata. In: Roda, C. (ed.): *Human Attention in Digital Environments*, Cambridge University Press (2011).
- [19] Pearson, K.: *On further methods of determining correlation*. Cambridge University Press, London (1907)
- [20] Niemann, K., Scheffel, M., Friedrich, M., Kirschenmann, U., Schmitz, H.-C., Wolpers, M.: Usage-based Object Similarity. In: Verbert, K. et al. (eds.): *Journal of Universal Computer Science*, special issue on Context-aware Recommender Systems (2010)
- [21] van Dongen, S.: *Graph Clustering by Flow Simulation*, PhD thesis, University of Utrecht (2000)
- [22] Kannan, R.: On Clusterings: Good, Bad and Spectral, 51(3), 497-515 (2004)