

Usage-based Clustering of Learning Objects for Recommendation

Marc-André Orthmann

Fachbereich Informatik

Hochschule Bonn-Rhein-Sieg

Grantham-Allee 20, 53757 Sankt Augustin, Germany

marc.orthmann@smail.inf.h-brs.de

Martin Friedrich, Uwe Kirschenmann, Katja

Niemann, Maren Scheffel, Hans-Christian

Schmitz, Martin Wolpers

Fraunhofer FIT

Schloss Birlinghoven, 53754 Sankt Augustin, Germany

{martin.friedrich, uwe.kirschenmann, katja.niemann

maren.scheffel, hans-christian.schmitz,

martin.wolpers}@fit.fraunhofer.de

Abstract— The growing amount of available information on the internet makes the process of filtering appropriate information an increasing challenge. Because currently existing approaches provide insufficient results in many cases, we propose a new way of relating objects based on their usage. We assume that objects which are significantly often used in the same session are semantically related. Thus, we build a usage-based relatedness graph, apply a graph-based clustering algorithm and evaluate the results with respect to semantic similarity measures. Our approach takes the learning domain into special consideration; its evaluation is performed within the Learning Object Repository MACE.

Keywords: text mining, clustering, usage data analysis, usage contexts, Contextualized Attention Metadata (CAM), learning object repositories, MACE

I. INTRODUCTION

The amount of information available on the internet is ever growing. This makes the process of filtering information an increasing challenge for the user. Instead of aimlessly browsing through large amounts of data, users fall back on using filtering or recommendation services [1]. Filtering systems remove unwanted information while recommender systems suggest to the users what they should read, do, watch or listen to. Especially in the area of Technology Enhanced Learning (TEL) where users work on specific topics it is important that they are supported to find documents which are semantically related to their learning goals or currently used items. For this reason, we explicitly do not consider Collaborative Filtering techniques because they sometimes tend to fail in finding semantically related documents [3].

To make use of automatic recommendation or advanced search services, a representation of the documents by semantic key features is required. To this end, most Learning Object Repositories (LORs) provide different kinds of metadata which are usually generated by manual creation or by automatic extraction of semantic features for each document. The automatic extraction of semantic features works well for text documents but produces insufficient results for other media types [10]. One way to avoid these drawbacks can be the use of social media technologies, in particular folksonomies. Social media technologies comprise metadata like ratings, tags or comments about resources which are created by a community. Especially tags provide

an effective way to represent user interests and help the user to find documents about a specific topic [4]. A disadvantage of using such social metadata is that they have to be added by a community and thus often contain ambiguous tags or synonyms. Furthermore, it is not assured that each tag is correct which can lead to wrong results.

Since both of the above mentioned approaches produce insufficient results we compensate these drawbacks by determining content relatedness of cross-media objects without relying on automatic semantic analysis or manual classification. We claim the hypothesis that the conjoint usage of data objects within a user session hints towards a semantic relatedness of these objects. We thereby 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 session then they are most probably related via the session task (e.g. learning math or history). This relatedness is a semantic/content-induced one, which e.g. means that both objects deal with a common superordinate concept [18]. We assume that the more often two objects are used together in different sessions, the stronger their usage-based and, thus, task-based connection is. Of course, we cannot assume that users are always fully focussed on one singular task: they take breaks and assess objects not related to their task. However, we expect that the usage-relatedness of semantically non-related objects will be rather weak and can be filtered out as noise in most cases.

For testing our hypothesis, we monitor the document usage within a LOR, namely the MACE repository, and build a usage graph by connecting objects which have been used together within the same session. Next, we calculate the relatedness of the objects based on using different metrics like the relative frequency with which the two objects occurred in the same session and adapt the usage graph accordingly. We apply a graph-based clustering method on the so constructed usage-based relatedness graph and evaluate the resulting clusters with respect to semantic relatedness measures.

Our approach is an exploratory one: we experimentally investigate the relation of usage and content, thereby taking some decisions ad hoc without proving their ‘necessity’ in advance. Our results, however, support that we are on the right track.

The rest of the paper is structured as follows. Chapter 2 introduces our approach of clustering objects based on their usage. Following our initial hypothesis we assume that the

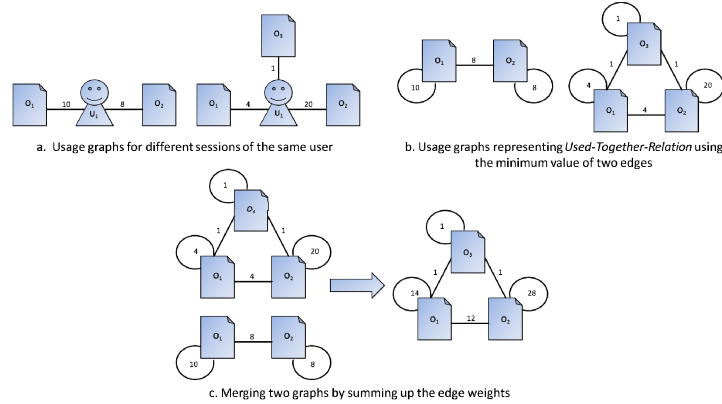


Figure 1. single steps to build the usage graph

objects within the usage-based clusters are semantically related. In chapter 3 we test this assumption, that is, evaluate our hypothesis. A summary is given in chapter 4.

II. COMPUTING USAGE RELATEDNESS

If our initial hypothesis that content-relatedness is positively correlated with usage-relatedness is true, then the sought-after procedure can be the following: we collect usage metadata by automatically observing user interactions with data objects, e.g. objects of a specific LOR. After building usage graphs by connecting documents that have been used together within the same session we define and calculate quantitative values of usage-relatedness and include these values in the graphs. Finally, we apply a clustering algorithm on the usage graphs and thereby generate classes of usage-related objects. According to our hypothesis, the members of these classes are not only related by usage but also by content. To test whether this is actually true, we evaluate the semantic similarities within the usage-based clusters.

A. Usage Graph

To represent the usage-based relatedness of objects we first built graphs containing all objects which were used by a specific user within one specific session: if user u_1 uses the objects o_1 and o_2 within the same session, then u_1 is related to both o_1 and o_2 . It might be that u_1 uses the objects more than once. To reflect this we weight the edges between u_1 and the objects by the number of respective usages. The graphs in Figure 1.a represent two sessions of user u_1 . In the first session u_1 used object o_1 ten times and object o_2 eight times; in the second session, she used o_1 four times, o_2 20 times and o_3 only once.

The first graphs relate users with objects. For directly relating objects, we removed the user nodes from these graphs but kept the – so far user-mediated – links between the objects. Moreover, we added a loop to each object node which is weighted by the frequency with which the object has been used within the session. The problem, still, is to determine plausible weights for direct links between objects. These edge weights can be determined in different ways, e.g. by considering the maximum, minimum or average value of the edges in the original graph. If the maximum value is

chosen, objects that are used only once become strongly connected to the other objects used in the session. If, for instance, o_3 has been used only once but another object o_2 has been used 20 times, then these objects are related by a weight of 20, which might be inappropriately high, especially when o_3 has been accessed accidentally. In order to alleviate this unwanted effect, we chose the minimum value from the original graph. We admit that this decision is rather ad hoc and might be revised in future experiments. (Cf. Figure 1.b.)

Building a graph for each user session can lead to huge numbers of graphs which is why we merged these graphs into one. Again, the edge weights of the resulting graph can be determined in different ways, e.g. by choosing the maximum, minimum or average value, or by summing up the values of the individual graphs. We decided to apply the sum of the edge weights because contrary to the other approaches this raises the weight for often used items while the weight of rarely used items stays low. This seems to be reasonable, when significant usage-based relations are to be discovered. (Cf. Figure 1.c.)

B. Relatedness Graph

Since the relatedness of two objects is not shown in this graph yet, we describe different relatedness measures which can be applied on the merged usage graph. In the following we will present three methods to calculate the relatedness between objects based on the usage graph. The generated relatedness values are then stored in a new graph which will serve as basis for object clustering. The generation of the relatedness graph based on different measures is illustrated in Figure 2.

1) *Relative Frequency:* The usage graph holds as edge weights the absolute frequencies of the joint occurrences of the objects represented by the corresponding nodes. One way to calculate the relatedness of two objects is to normalize these weights by considering the relative frequencies. The relative frequency of object o_1 and object o_2 , with $w(o_1, o_2)$ holding the weight of their connecting edge and $w(o_1)/w(o_2)$ holding the sum of all edges of o_1 and o_2 respectively, is calculated as follows:

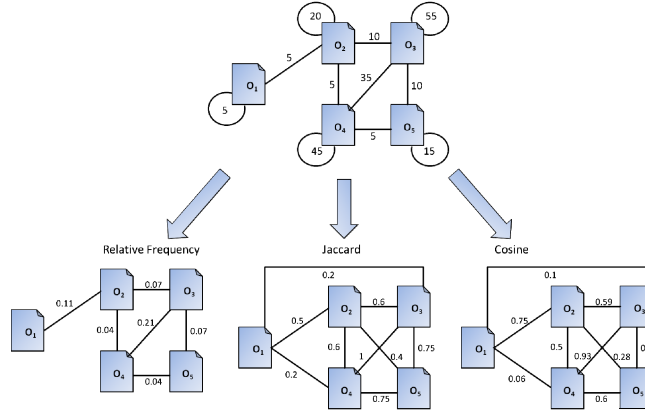


Figure 2. Graphs based on Relative Frequency, Jaccard Coefficient and Cosine Similarity

$$sim_{relFreq}(o_1, o_2) = \frac{w(o_1, o_2)}{w(o_1) + w(o_2) - w(o_1, o_2)}$$

The relative frequency is always a value between 0 and 1. Objects that were never used together get a relatedness value of 0.

2) *Jaccard Coefficient*: The Jaccard coefficient can be applied to two sets of objects and states the ratio of the cardinality of their intersection to the cardinality of their union [7]. Relating to the usage graph in Figure 2, an object o is represented by the set of objects which have been used together with o at least once in the same session. Given the sets O_1 and O_2 representing the objects o_1 and o_2 , the Jaccard coefficient is calculated as follows:

$$sim_{jaccard}(O_1, O_2) = \frac{|O_1 \cap O_2|}{|O_1 \cup O_2|}$$

The Jaccard Coefficient always has a value between 0 and 1. Contrary to the relative frequency, objects that were never used together can get a relatedness weight higher than 0 when their usage context – i.e. the set of objects they were used together with – are overlapping.

3) *Cosine Similarity*: Calculating the cosine similarity requires a vector representation for objects. Each node of the usage graph is represented as a vector with n dimensions according to the number of objects in the graph. The value of each dimension is defined by the weight of the edge between two objects. The cosine similarity now measures the relatedness between two objects by finding the cosine of the angle between their respective vector representations. Given the vectors V_1 and V_2 representing the objects o_1 and o_2 , the cosine similarity is calculated as follows:

$$sim_{cos}(V_1, V_2) = \frac{V_1 \cdot V_2}{\|V_1\| \cdot \|V_2\|}$$

Similar to the Jaccard Coefficient, the cosine similarity always returns a result between 0 and 1 and objects that were never used together can hold a degree of relatedness greater than 0 when they share a common usage context, i.e. they were used with the same objects.

C. Clustering: Iterative Conductance Cutting

To cluster the objects based on the usage graph we can apply different clustering algorithms. A common approach for graph-based clustering is Markov Clustering (MCL). The main idea of MCL is that a “random walk that visits a dense cluster will likely not leave the cluster until many of its vertices have been visited” [6]. Since the probability of paths with a higher length are more common within clusters than between different clusters, the probabilities associated with a pair of nodes lying in the same cluster will, in general, be relatively large as there are many ways of going from one to the other.

Since there is no guarantee that the MCL algorithm terminates we decided to apply Iterative Conductance Cutting (ICC) for clustering the graph which is similar to MCL regarding scalability [9]. The ICC 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 lesser importance to vertices which have few similar neighbours. The iteration ends when there are no more clusters which can be split without violating the threshold. An exhaustive search of the split minimizing the conductance would be exponential in time complexity [8]. There thus exists an approximation for applying ICC using Eigenvalues and Eigenvectors which hint on existing clusters.

III. EVALUATION OF THE USAGE-BASED CLUSTERING

In this section, we describe the evaluation of the approaches described in section II. Therefore, we introduce MACE as a test bed and describe how semantic similarities between MACE objects can be computed. These similarity values are then used as tentative gold standard to evaluate the usage-based clustering of the MACE objects.

A. Semantic Relatedness of MACE Objects

MACE (Metadata for Architectural Contents in Europe, [11]) is a European project that relates architectural learning

material stored in various repositories with each other to support students in finding relevant information. To enable connections between these learning objects, metadata representations of the objects are stored in the central MACE repository. The representations base on the MACE application profile which extends the Learning Object Metadata (LOM) standard [12] for architecture-specific needs.

The attributes stored in the LOM instances are used to build document vectors for all learning objects. To this end, the English titles and descriptions of the learning objects, the repositories the objects are derived from, their learning resource types as well as their free text tags, classifications and assigned competences are considered. The relatedness between two learning objects is then calculated using the cosine similarity measure between their representing document vectors. To get the distribution of the semantic relatedness of all the MACE objects, all possible pairs of objects have been calculated.

B. Usage-based Clustering of MACE Objects

While interacting with the MACE portal, users are monitored and their activities are recorded as CAM instances (Contextualized Attention Metadata, [13]). The CAM instances are analysed in order to construct a usage graph as described in section II.A. This usage graph is then transformed into three different relatedness graphs by applying the relative frequency measure, the Jaccard coefficient and the cosine similarity measure, respectively, as described in section II.B. Finally, the three graphs are taken as input for the ICC algorithm described in section II.C. As results, we get three different usage-based partitions of MACE objects. Based on the relative frequency measure, we cluster the 5960 considered MACE objects into 215 cluster sets; based on the Jaccard coefficient we cluster the MACE objects into 106 sets; and, based on the cosine similarity we cluster them into 67 sets.

C. Evaluation of the Usage-Based Clusters

The evaluation shall answer the question, whether a usage-based clustering of learning objects has implications for their semantic relatedness, i.e. whether learning objects in the same clusters do have a different relatedness than randomly drawn learning objects from the set of all learning objects. If this is the case, then the question arises whether the clustering improves or worsens the semantic relatedness. As such these two questions will be answered by applying two statistical methods, namely for the first the Kruskal-Wallis-Test and the second a test for the t-distribution.

To test whether an overall effect of the usage-based clustering is detectable at all, an ANOVA could be the first choice but was dismissed as the required pre-conditions were not met (semantic relatedness is not normally distributed as a Kolmogorov-Smirnov-Test revealed [17], homogeneity of variances within clusters is not given). Therefore, the non-parametric alternative for the ANOVA, the Kruskal-Wallis-Test was chosen as it just relies on ranked data and does not make assumptions on an underlying normal distribution or homogeneity of variances.

For testing whether the clustering leads to an improvement of the semantic relatedness, it can reasonably be assumed that the means of the relatedness values within clusters t-distribute around the overall mean-value of the population of MACE objects [15]. As the population is given, it is possible to check whether the different clusters deviate significantly from the population of all learning objects without committing an accumulated alpha-error that would bias the results by chance.

1) *Kruskall-Wallis-Test*: The Kruskal-Wallis-Test [16] evaluates whether the medians of certain groups' ranked data (dependent variable) differ systematically or not. Therefore, several tests had to be done that distinguish between the entire set of MACE objects and the cluster sets of the (i) relative frequency-based clustering, (ii) Jaccard coefficient-based clustering and (iii) cosine similarity-based clustering.

For each usage-based relatedness measure (relative frequency, Jaccard coefficient, cosine similarity) it is tested whether the semantic relatedness values of the respective clustering – all partition sets taken together – differ significantly from the overall median of relatedness of the entire MACE set. To this end, a null-hypothesis (H0) is formulated:

H0: The learning objects are taken from the same population. Usage-based clustering does not have an effect on semantic relatedness.

Cluster sizes were determined indirectly by the threshold of the ICC algorithm. A conductance value of 0.25 leads to an average cluster size of 33.25.

TABLE I. KRUSKALL-WALLIS-TEST

<i>Similarity Measure</i>	<i>Chi-squared</i>	<i>df</i>	<i>p</i>
Relative Frequency	424713.3	214	<0.001
Jaccard Coefficient	23827.08	105	<0.001
Cosine Similarity	17394.29	66	<0.001

Table 1 displays the results of the analysis. The results show that for all three usage-based relatedness measures, clustering has a highly significant impact on mean semantic relatedness. That is, the null hypothesis H0 has to be dismissed, and the competing H1 is strongly supported.

H1: The learning objects are not taken from the same population. Usage-based clustering has an effect on semantic relatedness.

The results are to be taken as a first hint not as a final analysis. It still has to be checked whether the clustering leads to an improvement or not.

2) *Test for t-Distribution*: The Kruskal-Wallis-Test checks whether the entire clustering has an effect on semantic relatedness. However, it does not compare the individual cluster sets with the overall set of all MACE

objects. As in our case the population of all MACE objects is available, no additional post-hoc tests must be applied. For comparing individual cluster sets with the overall set, it can simply be tested whether the cluster means of semantic relatedness differ significantly from the overall mean of the population. To this end, the null hypothesis for each single cluster set is formulated:

H0: The mean semantic relatedness of a specific cluster set does not differ from the mean semantic relatedness of the whole population of learning objects.

Accordingly, H1 states:

H1: The mean semantic relatedness of a specific cluster set systematically differs from the mean semantic relatedness of the whole population.

Each usage-based relatedness measure leads to a usage-based clustering. The t-test to evaluate H0/1 was applied to every cluster set of each clustering. Again, H0 is to be dismissed by a significance level of $\alpha=0.05$. The effect of getting a significant cluster set by chance (accumulation) was not accounted for as the sets were of different sizes.

TABLE II. T-DISTRIBUTION-TEST

Similarity Measure	Number of Cluster Sets	P++ ^a	P-- ^b
Relative Frequency	215	64,65%	7,44%
Jaccard Coefficient	106	78,30%	6,60%
Cosine Similarity	67	68,66%	13,43%

a. Percentage of clusters significantly over-average ($p<0.05$)

b. Percentage of clusters significantly under-average ($p<0.05$)

Table 2 shows the results of the evaluation: irrespectively of the relatedness measure leading to a clustering, the semantic density of the clear majority of cluster sets is significantly higher than the semantic density of the overall population (where semantic density is defined as the mean value of semantic relatedness). Especially the Jaccard coefficient provides good results. If one randomly chooses one cluster set, then the chance that the included learning objects show a higher similarity than at a random draw is about 77%. This is a very good result supporting our initial hypothesis. Both the Kruskal-Wallis-Test and the test for the t-distribution support our initial hypothesis that usage-relatedness as defined here correlates with semantic relatedness.

IV. CONCLUSION

The support of users in finding appropriate documents related to their current context requires a semantic representation of these documents. Currently used approaches like automatic extraction or manual creation of semantic features lead to efficiency and precision problems. To solve these problems new approaches are required which circumvent the known shortcomings. In this paper, we

explored a new way to relate learning objects that leaves aside semantic information. We tested this approach within the MACE LOR by evaluating about 5960 resources. The results very are promising. We assume that our approach can be fruitfully used for recommendation and filtering systems. The proof of this assumption is subject to further investigation.

ACKNOWLEDGMENT

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

REFERENCES

- [1] Niemann, K., Scheffel, M., Friedrich, M., Kirschenmann, U., Schmitz, H.-C., Wolpers, M.: Usage-based Object Recommendation. In: Journal of Universal Computer Science, Vol. 16, No. 16, 2010, pp. 2272-2290.
- [2] van Metern, R. & van Someren, M.: Using Content-Based Filtering for Recommendation, Technical report, Foundation for Research and Technology - Hellas (2002)
- [3] Smeulders, A. W., Member, S., Worring, M., & Santini, S.: Content-Based Image Retrieval at the End of the Early Years. *Analysis*, 22(12), 1349-1380 (2000).
- [4] Zhao, Y. E.: Tag-based Social Interest Discovery. *Social Networks*, 675-684 (2008)
- [5] Duan, M., Ulges, A., & Breuel, T. M.: Style Modelling for Tagging Personal Photo Collections. *Style (DeKalb, IL)*.
- [6] van Dongen, S.: Graph Clustering by Flow Simulation, PhD thesis, University of Utrecht (2000)
- [7] Matsumoto, Y.: Lexical Knowledge Acquisition. In: Mitkov, R.: The Oxford Handbook of Computational Linguistics. Oxford University Press, Oxford, 2003, 395-413.
- [8] Foggia, P., Percannella, G., Sansone, C., Vento, M., Elettrica, I., Salerno, U.: A Graph-Based Clustering Method and Its Applications, 277 – 287 (2007)
- [9] Kannan, R.: On Clusterings: Good, Bad and Spectral, *SI(3)*, 497-515 (2004)
- [10] Greiner, F.: Benutzerorientierte Evaluation von Content Based Image Retrieval-Systemen mit automatischer Beschlagwortung, Diplomarbeit, Universität Regensburg (2009)
- [11] Stefaner, M.; Vecchia, E. D.; Condotta, M.; Wolpers, M.; Specht, M.; Apelt, S. & Duval, E.: MACE - Enriching Architectural Learning Objects for Experience Multiplication., in Erik Duval; Ralf Klamma & Martin Wolpers, ed., 'EC-TEL', Springer, , pp. 322-336 (2007)
- [12] LOM IEEE Standard for Learning Object Metadata, IEEE Std 1484.12.1, 2002.
- [13] Wolpers, M., Najjar, J., Verbert, K., Duval, E.: Tracking Actual Usage: the Attention Metadata Approach. *Educational Technology & Society* 10 (3), 106-121. (2007)
- [14] Porter, M.-F.: An algorithm for suffix stripping. *Program* 14: 130-137. (1980)
- [15] Bortz J. Statistik für Human- und Sozialwissenschaftler, 6th edn. Heidelberg: Springer Medizin Verlag 2005.
- [16] Sawilowsky, S. and Fahoome, G. 2005. Kruskal–Wallis Test. *Encyclopedia of Statistics in Behavioral Science*.
- [17] Massey F. J.: The Kolmogorov-Smirnov Test for Goodness of Fit, *Journal of the American Statistical Association*, Vol. 46, No. 253. (1951), pp. 68-78.
- [18] Hoey, M.; Corpus linguistics and word meaning. In: Lüdeling, A., Kytö, M. (eds.): *Corpus Linguistics. An International Handbook. Volume 2*, de Gruyter, Berlin