Object Recommendation based on Usage Context

Martin Friedrich, Katja Niemann, Maren Scheffel, Hans-Christian Schmitz, Martin Wolpers
Fraunhofer Institute for Applied Information Technology FIT, Schloss Birlinghoven,
53754 Sankt Augustin, Germany
{martin.friedrich, katja.niemann, maren.scheffel,
hans-christian.schmitz, martin.wolpers}@fit.fraunhofer.de

1 Introduction

With more and more information available online it can be difficult to find what one is really looking for, what is interesting in the current circumstances or what is best to be looked at next. Instead of aimlessly browsing through large amounts of data by themselves, people often use recommender systems [01] to suggest to them what to read, do, buy, watch or listen to next. In this paper we will claim the hypothesis that usage similarity gives rise to content similarity and can thus be used for recommendations. We will therefore define the notion of a usage context profile (UCP) for data objects. The UCP of an object *o* can be derived from a set of usage histories; it contains the objects that were used before and after *o* was accessed. We will then introduce a similarity measure for UCPs. We will argue that recommendation can be improved (i) by comparing the current usage history of a user with UCPs of data objects and (ii) by relating data objects according to their usage similarity.

We refer to the MACE project[1] as a test bed. MACE (*Metadata for Architectural Contents in Europe*, [02]) is a European project where digital learning resources about architecture, stored in various repositories, are related with each other across repository boundaries to enable powerful ways of finding relevant information. While interacting with the MACE portal, users are monitored and their activities are recorded as CAM (*Contextualized Attention Metadata*, [03]). Activities include search, access and metadata provision activities like tagging and rating. We use the CAM recordings to derive UCPs of respective MACE objects. In the long term, we aim to improve the system by recommending objects to the user for his current context based on his usage history and UCPs of MACE objects.

2 Recommender Systems

The three approaches of recommender systems commonly implemented are collaborative filtering, content-based filtering and hybrid filtering which combines aspects of the other two approaches [04]. Content-based systems match item descriptions to the user's interests in her profile. User profiles can be built explicitly by asking users about their interests or implicitly by a user's given ratings. However, an item's content is not always available making it difficult to match with a profile. User profiles in turn first have to evolve to suffice for recommendations and cannot really mirror the current interest but only an average short term or average long term one. For users with many interests (or ratings or purchases) though, the algorithm often has to use a subset of the interests to scale and perform well, resulting in too general or too narrow recommendations [05].

Systems based on collaborative filtering make use of user ratings on items, allowing user-based as well as item-based collaborative filtering approaches. In user-based approaches, an item is suggested to a user based on ratings of that item by users most similar to her, making it

[1] http://www.mace-project.eu

necessary to previously calculate similarities between users. This method is computationally very expensive. Reducing the data size might lead to better scaling but usually worsens the recommendation quality [05]. If a user is only compared to a small group of other users, those might not be similar enough. If only a smaller range of items is considered, recommendations would be restricted to certain topics. Thirdly, if very popular or very unpopular items are not considered, they will hardly be looked at.

For item-based approaches the similarity between items has to be calculated and only then can the recommendation of an object be based on a user's ratings of similar items. As this can be done offline, such algorithms are quicker and scale more easily [06]. The similarity of two items depends on how often they both have been used/bought/looked at by different users, either during one session or in general. If a user for example just used object o_1, she might get the recommendation to use object o_2 because that was the object mostly used by others who also used o_1.

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 usage contexts have to be similar. We can therefore recommend object o_2 to a user who previously used o_1 based on the fact that o_1 and o_2 have similar usage contexts which not necessarily entails that they have both been used by the same users.

3 Similarity Calculation

Within this section we introduce three different types of similarities and their calculations. We start with context similarity which is based on the usage contexts of objects. The metadata similarity is based on different attributes from the learning objects' metadata. Finally, we manually compare a selection of learning objects based on their content.

3.1 Context Similarity

For every object used we create a usage context profile (UCP) which comprises at least one usage context, one for every session in which the object was used. We define a usage context of a data object *o* as consisting of a pre-context and a post-contest. The pre-context is the sequence of objects that were accessed before *o*, and the post-context is the sequence on objects that were accessed after *o*, within the same session. A MACE-session consists of all recorded actions between a user's log in and log out event. An object can be used several times and in different contexts. The UCP of this object is the set of its different usage contexts. Within MACE, UCPs can be derived from by mere data transformation and integration of the CAM recordings.

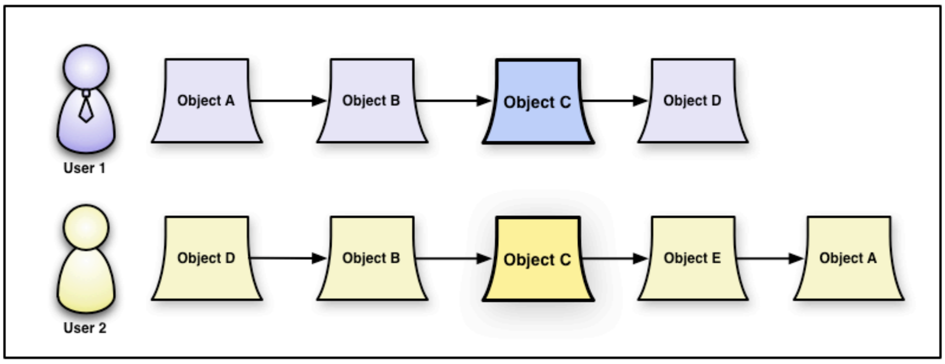

Figure 1: learning paths of two different users

More detailed: for every object, its UCP consists of one or several usage contexts, that is, pairs of pre- and post-contexts: $\{\langle \mathrm{C}_1^{\mathrm{pre}}, \mathrm{C}_1^{\mathrm{post}}\rangle, \ldots, \langle \mathrm{C}_n^{\mathrm{pre}}, \mathrm{C}_n^{\mathrm{post}}\rangle\}$. Pre- and post-contexts are lists of objects. Figure 1 shows two sessions of two users. Based on the histories of these sessions, a UCP for object C can be generated. It is a set of two contexts consisting of the objects that were used before and after C:

$$\{\langle (ObjectA, ObjectB), (ObjectD)\rangle, \langle (ObjectD, ObjectB), (ObjectE, ObjectA)\rangle\}$$

A UCP is a set of usage contexts. Similarity of two contexts arises from similarities of the associated pre- and post-contexts. Similarities of pre- and post-contexts are calculated using the Jaccard similarity measure [07] where C_1 and C_2 are pre- or post-contexts:

$$(1)\ \ simPCon(C_1, C_2) = \frac{|\cap(C_1, C_2)|}{|\cup(C_1, C_2)|}$$

The similarity of two contexts – pre-/post-context pairs – is defined as the arithmetic mean of the pre- and the post-context similarities:

$$(2)\ \ simCon\left(\langle \mathrm{C}_1^{\mathrm{pre}}, \mathrm{C}_1^{\mathrm{post}}\rangle, \langle \mathrm{C}_2^{\mathrm{pre}}, \mathrm{C}_2^{\mathrm{post}}\rangle\right) = arithMean\left(simPCon\left(\mathrm{C}_1^{\mathrm{pre}}, \mathrm{C}_2^{\mathrm{pre}}\right), simPCon\left(\mathrm{C}_1^{\mathrm{post}}, \mathrm{C}_2^{\mathrm{post}}\right)\right)$$

The similarity of two UCPs UCP_1 and UCP_2 is defined as the arithmetic mean of the summarised pair-wise usage context similarities:

$$(3)\ \ simUCP(UCP_1, UCP_2) = \frac{\sum_{\langle X,Y\rangle \in UCP_1 \times UCP_2} simCon(X,Y)}{|UCP_1 \times UCP_2|}$$

3.2 Metadata Similarity

In the MACE project, only metadata representations of learning objects are stored on a central server. The representations base on the MACE application profile that is based on the Learning Object Metadata (LOM) standard [08]. The MACE application profile comprises several categories which are used to specify the learning object in more detail, such as the general category where basic information about the learning object is stored and the annotation category where comments about the usefulness for education of the learning object and the origins of this comments can be stored.

To calculate the similarity of two learning objects based on their metadata, we consider the following assortment of the available information about the learning object: repository, titles, descriptions, learning resource types, user tags, classifications and competencies of the learning object. Each learning object belongs to exactly one repository where it is stored. It holds one or more titles and descriptions, which are signed with a language tag. To make the learning objects more comparable, only English titles and descriptions are taken into account. 4.5% of the about 5000 considered learning objects are not described in English and were thus excluded from further analysis. Each learning object is assigned with the learning resource types it comprises, e.g. a text containing a figure is assigned with the learning resource types ‘narrative text’ and ‘figure’. These terms are from a controlled vocabulary and are therefore comparable. Repository name, titles, descriptions and learning resource types are derived directly from the repository.

However, MACE offers also the possibility to users and domain experts to edit parts of the metadata. Tags are free text and can be assigned to learning objects by logged in users. 52% of the considered learning objects are already tagged this way. Classifications and competencies are each defined in a controlled vocabulary and can only be set by domain experts. The classification vocabulary is a taxonomy consisting of 2884 terms. The classification terms are used to describe the learning objects in more detail and to simplify the search for learning objects. 47.9% of the considered learning objects are already assigned with classifications, where each of them contains an average of 3.1 classification terms. The competency vocabulary contains 107 terms which are used to describe the suitability of learning objects for the obtainment of special competencies, e.g. ‘Knowledge of internal environment control’ and

‘Understanding interaction between technical and environmental issues’. 38.5% of the considered learning objects are already assigned with competencies, where each of them contains an average of 5.8 competency terms.

Before being used for similarity calculation, the titles and descriptions are pre-processed. In the first step, stop words are removed. In the second step a stemming using the Snowball Stemmer [09] is carried out. To compare the learning objects using the free text attributes (titles, descriptions, tags) document vectors describing the learning objects are generated and the similarity is calculated using the cosine distance which measures the similarity between two vectors by calculating the cosine of the angle between them:

$$(4)\quad cos(\theta) = \frac{X \cdot Y}{\|X\| \|Y\|}$$

where X is the vector of object o_1 and Y the vector of object o_2. To compute the similarity of learning objects using the attributes which are described through controlled vocabularies (classifications, competencies, learning resource types, repository), the Jaccard Coefficient, as described in 3.1, is used. The computed similarities for two learning objects can be combined in several ways that will be outlined in more detail in section 4.

3.3 Content Similarity

Since metadata are limited in that they do not fully represent a learning object’s content, we manually compared those 100 learning object pairs with the highest usage context similarity. Since the manual proof for content similarity does not produce an explicit similarity value that can be compared to other content similarity values, we focused on finding the content overlap of two learning objects. Therefore, we have accessed the content of the learning object pairs directly and compared them with each other. In the following, we give an overview of different categories of content similarity we discovered which are not entailed in the metadata.

Many of the checked learning object pairs show text documents which handle the same topic such as ‘risk factor analysis’, ‘low energy construction’, ‘music and architecture’ or ‘fire safety’. We have also identified learning object pairs where one of them was an exercise and the other one showed a text about the same topic. In one case we found a pair referencing websites about graphical algorithms. One of those websites provides the opportunity to browse and see examples of shape generating algorithms (including pictures), while the other one provides the possibility to create and test such kind of algorithms. Some of the learning object pairs refer to different web pages of the same domain. For instance, both objects refer to different entries of the Archiplanet Wiki[2] and show similar buildings. The similarity of these buildings is expressed by different attributes like similar construction date, architectural style, building type (e.g. commercial buildings like banks) or the construction system containing the building material. Even though the MACE application profile offers the possibility to store such information, they were not contained in the metadata.

Another category of content similarity is about picture similarity. Thus we discovered documents showing similar pictures, e.g. photos, sketches or models of the same building or construction activity like panel cladding. Furthermore, we could also identify learning object pairs containing a geographical similarity of the displayed content, e.g. pairs which represent websites containing pictures or articles of different historical buildings in the same town.

4 Analysis

We pick up intuitions from linguistics and language processing (namely from the notion of a ‘paradigmatic relation’, [10] and [11]) and claim the hypothesis that context similarities give

[2] http://www.archiplanet.org/wiki/Main_Page

rise to content similarities. To test this hypothesis we analyse learning objects inside the federated MACE repository.

4.1 MACE as Test Bed

At the time of our data collection there had been 1686 different users within the MACE test bed. 430 of these users had registered MACE accounts, the other 1256 users logged in as guests. We identified 4396 sessions – 3130 sessions of registered users and 1266 sessions of guests. In total, 13525 learning objects were accessed: 84.5% of them were accessed by registered users only, 5.6% by guests only and 9.9% by both groups. As the number of objects accessed by guests was very low and their average session length was only 2.34 accessed learning objects, we only analysed the sessions of registered users which had an average length of 13.67 learning objects per session. Every learning object occurred on average in 3.35 sessions. Due to performance issues we only considered those learning objects with more than two accesses (about 5000). Table 1 shows the distribution of context similarity greater than 0.5 for learning object pairs: 698 learning object pairs have a context similarity between 0.5 and 0.55 opposed to seven pairs with a context similarity between 0.95 and 1.0.

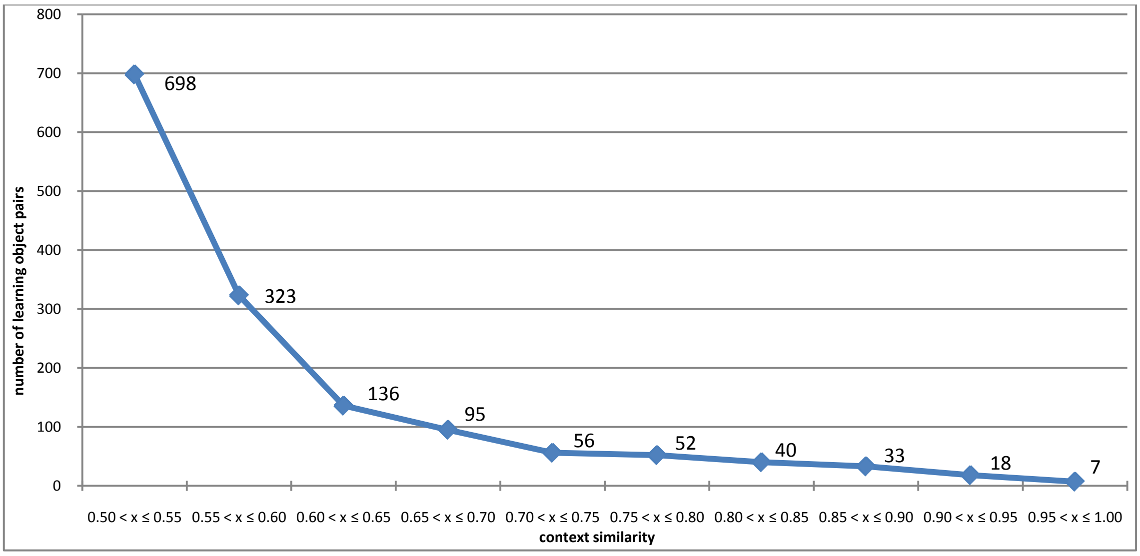

Table 1: context similarity distribution for context similarity > 0.5

4.2 Results

To prove our hypothesis, we used the metadata of the learning objects as shallow content representations and calculated the Pearson’s Correlation Coefficient [12] between metadata and context similarity with the following formula:

$$(5)\quad r_{yx} = \frac{cov(y,x)}{\sqrt{var(y)*var(x)}}$$

We calculated the metadata similarity for every attribute on its own and for all possible combinations. The correlation between single feature metadata similarities and context similarities always ranges between 0.2 and 0.3. When combining some attributes, e.g. all free text metadata (titles, descriptions and tags), the correlation rises up to 0.35. However, the highest correlation result, 0.42, is achieved when all metadata features are taken into account.

Although this is a significant correlation due to our large sample size of about 12 million learning object pairs, it is not a proof but an indication of our hypothesis.

In order to further investigate this indication and to circumvent the fact that the metadata only offers a shallow content representation, we manually compared the 100 learning object pairs with the highest usage context similarity. We found that 92% of the considered learning object pairs showed similarities as described in 3.3 above, 4% were not accessible due to permission rights and only 4% showed no similarity at all. These content similarities are in many cases not entailed in the learning object's metadata but can be detected by the user due to her prior knowledge (e.g. knowing that two buildings were designed by the same architect). As this shows that usage similarities can hint at content similarities that have not been considered so far, we see it as a motivation to continue on this track.

4.3 Usage Context Profiles for Recommendations

User context profiles (UCPs) can be used to recommend objects with different methods. One approach is to compare a user's current usage history – that is, the sequence of objects she has accessed so far – with pre-contexts of all other learning objects and recommend those objects that have highly similar pre-contexts. Another approach is to recommend learning objects that have similar UCPs – i.e. pre- and post-contexts – to the UCP of the object currently in use. A third approach is to also generate UCPs for search query terms and use them for the ranking of search results. When a user enters a query term, the pre-contexts of that query term are compared to the user's current usage history. Objects from the query term's post-contexts whose corresponding pre-contexts are similar to the user's usage history are ranked higher in the list of search results.

5 Conclusion and Future Work

In this paper we introduce a new idea for recommending learning objects based on their usage context. The hypothesis behind our approach that usage context similarity is an indication of content similarity is supported by our results. This motivates us to further develop this approach. We plan to include our findings into the MACE project and thus enable a large scale evaluation.

As we only considered part of the available data (i.e. objects with more than two accesses), an important aim is to improve the scalability of the usage context similarity calculation. We also aim at improving its quality. The length of usage contexts for example could influence the context similarity. A pair sharing 3 of 4 occurring objects could be deemed less similar than a pair sharing 6 out of 8 objects although they both have a similarity of 0.75. Longer usage contexts could therefore be weighted higher. Additionally the order of the learning objects could also be taken into account when calculating the context similarities.

6 References

[01] 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), 2005, pp. 734-749.

[02] 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, 2007, pp. 322-336.

[03] Wolpers, M., Najjar, J., Verbert, K., Duval, E.: ‘Tracking Actual Usage: the Attention Metadata Approach’. *Educational Technology & Society* 10 (3), 2007, pp. 106-121.

[04] Candillier, L., Jack, K., Fessant, F., Meyer, F.: ‘State-of-the-Art Recommender Systems’. In: Chevalier, M., Julien, C., Soule-Dupuy, C. (eds.): *Collaborative and Social Information Retrieval and Access – Techniques for Improved User Modeling*. Idea Group Publishing, 2009, pp. 1-22.

[05] Linden, G., Smith, B., York, J.: ‘Amazon.com Recommendations – Item-to-Item Collaborative Filtering’. *IEEE Internet Computing* 7 (1), 2003, pp. 76-80.

[06] Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: ‘Item-Based Collaborative Filtering Recommendation Algorithms’. *Proceedings of the 10th International World Wide Web Conference*, ACM Press, 2001, pp. 285-295.

[07] Matsumoto, Y.: ‘Lexical Knowledge Acquisition’. In: Mitkov, R.: *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford, 2003, pp. 395-413

[08] *LOM IEEE Standard for Learning Object Metadata*, IEEE Std 1484.12.1™, 2002.

[09] Porter, M.-F.: ‘An algorithm for suffix stripping’. *Program* 14: 1980, pp. 130-137.

[10] Saussure, F. de: *Course in General Linguistics*. Open Court Publishing, Chicago (re-print 1986).

[11] Heyer, G., Quasthof, U., Wittg, T.: *Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse*. W3L-Verlag, Herdecke, 2006.

[12] Pearson, K.: *On further methods of determining correlation*. Cambridge University Press, London, 1907.