Getting a grasp on tag collections by visualising tag clusters based on higher-order co-occurrences

Getting a grasp on tag collections by visualising tag clusters based on higher-order co-occurrences

Katja Niemann, Sarah León Rojas, Martin Wolpers
Fraunhofer FIT
Schloss Birlinghoven
53754 Sankt Augustin, Germany
{katja.niemann, sarah.leon.rojas, martin.wolpers}@fit.fraunhofer.de

Maren Scheffel, Hendrik Drachsler, Marcus Specht
Open University of the Netherlands
Valkenburgerweg 177
6419 AT Heerlen, The Netherlands
{maren.scheffel, hendrik.drachsler, marcus.specht}@ou.nl

ABSTRACT
Tagging learning resources in repositories or web portals offers a way to meaningfully describe these resources. The more tags there are, however, the more difficult it is to find one's way around the repository, especially when they are user-generated free-text tags. This paper therefore presents a visualisation of tag clusters based on higher-order co-occurrences that allows users of such repositories a plain but simple way of exploring them in an intuitive manner.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering, Information filtering, Search process, Selection process*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing; J.1 [**Administrative Data Processing**]: —*Education*; K.3.1 [**Computers and Education**]: Computer Uses in Education

General Terms
Algorithms, Experimentation, Language

Keywords
clustering, higher-order co-occurrences, tags, technology enhanced learning, visualisation

1. INTRODUCTION
Many educational web portals allow users to manually enrich the offered learning resources with social metadata like comments and free-text tags. It has been shown that tags in particular provide powerful knowledge that can be used to improve the quality of searching and recommendations [4, 7]. Similar to automatically extracted keywords, tags thus offer a way to get a quick grasp on the content or theme of multimedia objects. Especially when dealing with multimedia objects that provide little or no textual context (e.g. photos or videos) tags provide meaningful descriptors of these objects [8].

A common problem, however, when relying on tags is that they are often user-generated and not restricted to a closed vocabulary. Different users can tag the same learning resource with different tags leading to a large collection of rarely used but highly related tags. The use of singular or plural versions of the same word, the same word in different languages or different words with the same meaning, i.e. synonyms, can also lead to problems when relying on tags in order to get an overview on a collection of learning resources. In order to detect unknown relations between tags they therefore need to be contextualised.

Based on an approach of visualising large document collections according to the documents' keywords [6] we suggest to use a visualisation of tag relations that allows users to quickly get a grasp of the resources offered by a learning portal and to dig deeper to get an understanding of certain subject areas. Instead of clustering the learning objects according to their content, however, we cluster the tags according to their higher-order co-occurrences and then present them in a clearly arranged and intuitive manner. The creation of higher-order co-occurrences is a well-known approach in corpus linguistics to discover semantic relations between words based on their usage in text documents [2]. We adapt this approach by analysing the assignments of tags to learning resources instead of the occurrences of terms in sentences or text documents.

The paper is structured as follows. Chapter 2 gives a short overview of related work. Chapter 3 describes the approach of higher-order co-occurrence clustering to group tags with similar meanings, followed by the description of the MACE data set in chapter 4 which is used in this paper. Thereafter, chapter 5 describes the visualisation of the tag clusters and chapter 6 discusses the results. Finally, chapter 7 holds a conclusion and an outlook on future work.

2. RELATED WORK
According to Rivadeneira et al. [5], a meaningful visualisation of tags supports four main functions: (1) search, i.e. tags can be directly included in the search process and, thus, enhance the findability of items, (2) browsing, i.e. the visualisation offers a central entry point for users that know what

they are looking for but not what exactly to search for, (3) impression formation / gisting, i.e. the visualisation allows users to get a quick grasp on the items' subject areas, and (4) recognition, i.e. the users are offered the possibility to understand different aspects of certain information.

The most common approach to visualise a large number of tags is the creation of tag clouds. Here, the relative size of each tag stands in relation to its frequency in the tag collection. Nowadays, many tools are available that allow an easy integration of personalised tag clouds in web sites, e.g. TagCrowd[1] and Wordle[2]. While there is a huge potential inherent in tag clouds they also suffer from some issues, e.g. the missing semantic between the visualised tags [1, 9]. In order to deal with this, tag clouds have been created that analyse (first-order) co-occurrences between the tags and group tags that often co-occur [11]. Here, *similar* tags do not necessarily reference to the same semantic concept but are linked by the resources they have in common [3]. Another problem of tag clouds is that many frequent tags often dominate the whole tag cloud and less frequent tags and their concepts get lost [1].

This paper presents a clustering approach for tags that is based on higher-order co-occurrences, i.e. a corpus linguistic technique to find semantically related terms [2]. This way we aim to discover and visually cover all subject areas even though it might not be possible to display all single tags.

3. HIGHER-ORDER CO-OCCURRENCE CLUSTERING OF TAGS

The creation of higher order co-occurrences is a corpus-linguistic approach to exploit the usage context of linguistic entities in order to find semantic relations. Two linguistic entities are defined to be co-occurrences if they occur in at least one common usage context, e.g. in a sentence. For example, the word *dog* often co-occurs with the words *bark*, *growl*, and *sniff* among others.

In order to calculate the significance of a co-occurrence statistical association measures are used. Thereafter, the most significant co-occurrences must be selected for each term. Since there is no standard scale of measurement to draw a clear distinction between significant and non-significant occurrences, there are two ways to do so, i.e. by selecting only the n most significant co-occurrences for each resource or by using a threshold.

The significant co-occurrences of an entity form its first-order co-occurrence class and entities which co-occur in first-order co-occurrence classes are second-order co-occurrences. These second-order co-occurrence classes again can be used as input to calculate third order co-occurrences and so forth. When this procedure is repeated several times, the higher-order co-occurrence classes tend to get stable, i.e. their elements do not change any more. This indicates that there exist universal relations between the entities in the remaining classes that induce their aggregation again in each iteration step. In fact, these stable higher-order co-occurrence classes have shown to usually hold semantically related entities.

[1]http://tagcrowd.com/

[2]http://wordle.net/

Heyer et al. [2] show this for the co-occurrences of *IBM*, among other words. Their investigations are based on text corpora collected for the portal wortschatz.uni-leipzig.de, the German treasury of words. The first co-occurrence class is rather heterogeneous, and contains words like *computer manufacturer*, *stock exchange*, *global* and so on. After some iterations of computing higher-order co-occurrence classes, however, the classes become more homogenous and stable. The tenth order co-occurrence class only contains names of other computer-related companies like *Microsoft*, *Sony* etc.

In the given scenario we do not have sentences in which the tags occur. However, the tags are assigned to learning resources which can be considered to represent usage contexts. Thus, two tags are co-occurrences if they are assigned to at least one common learning resource. In order to calculate the significance of two tags, the association measure Mutual Information (MI) is used which compares the observed frequency O of a co-occurrence with its expected frequency E, see formula 1.

$$\mathrm{MI} = log_2 \frac{O}{E} \qquad (1)$$

Here, selecting the n most significant co-occurrences for each tag would imply to have a pre-defined cluster size which is not desirable, thus, a threshold is used. Because the calculated significance scores for resource pairs are only comparable if they have one resource in common, a resource-specific threshold is used to distinguish between relevant and non-relevant co-occurrences. Here, this threshold is calculated for each learning resource by averaging the significance values of all its co-occurrences and multiplying the result with a regulation constant α which has a value of 0.95 in the presented experiment.

4. THE MACE DATA SET

The MACE[3] (Metadata for Architectural Contents in Europe) project relates digital learning resources about architecture with each other across repository boundaries to enable a simplified discovery and access [10]. Users are able to search for learning resources and filter the results, e.g. according to their language, the original repository, and the classification terms they hold. Furthermore, the portal offers a social search based on tags, a location search based on the geographical coordinates of buildings represented through learning resources, and a competence search based on the competencies the learning resources aim to impart. Registered and logged-in users are able to rate, tag, and comment on learning resources. Additionally, they can follow the metadata provision activities of other users.

The MACE data set holds 117,907 events on 12,442 learning resources conducted by 630 registered users. 70.8% of the learning resources hold tags in which each tagged learning resource holds on average 6.59 tags. Overall, the users assigned 13,291 distinct tags of which 73% are only used once and only about 4% of the tags are added to more than 10 learning resources.

[3]http://mace-project.eu/

5. VISUALISATION

When creating a visualisation of the tag clusters for the MACE data set we decided to not present tags in the visualisation that are assigned to only one or two learning resources. Only clusters that hold more than five tags are selected for presentation. Finally, the two most frequent tags are selected as title for each cluster. If a cluster's most frequent tags significantly overlap, the less frequent one is neglected and the next frequent tag is selected.

After this data processing, the tag clusters and all attached information are written to a JSON file. The visualisation is realised using the Data-Driven Documents D3.js framework[4], i.e. a JavaScript library, paired with HTML, CSS and JQuery to process the previously created JSON files.

Figure 1 shows the default starting view of the visualisation[5]. The tag clusters are represented by circles and are ordered according to their size in the form of a spiral with the largest cluster having the largest circle and being positioned at the outside of the spiral and the smallest cluster being in the middle of the spiral. Here, the size of a tag cluster depends on the number of learning objects that are referenced by the tags belonging to it. Additionally to size and position, every cluster has its own color and is labelled with its two most representing tags to enable the users to quickly get a grasp on the clusters' content.

By clicking on a cluster, the view changes and the visualisation zooms into to the chosen cluster for which up to 20 tags become visible. We chose this number to not overload the visualisation. In order to continue the circle approach used for the clusters, we adapted the common usage of font size, coloring and word positioning in tag clouds and used sized and spirally ordered circles for the tags as well. On the right side of the visualisation, a list of all the learning resources that are associated with that cluster is given showing the resources' title, media type, and language additionally to the list of all tags assigned to it. All resource titles link to the original resource.

Clicking on a tag circle results in a new list next to the visualisation in which all resources assigned with that tag are given. By clicking on a specific tag, its circle is highlighted and the object list only displays those resources that are assigned with the highlighted tag, see figure 2.

6. DISCUSSION

This chapter provides an insight on the eleven clusters shown in the visualisation, discusses the topics they cover including their relations, and reference further distinctive features. The following cluster descriptions are ordered by the size of the clusters, i.e. from the outside of the spiral to its center. Whenever needed, the tags' English translations are given in brackets. Here, if two tags hold the same English translation it is only given once.

Cluster 1: cubierta / aislante (cover and insulation). This cluster's tags, which are mainly in Spanish, name meth-

[4]http://d3js.org/

[5]The visualisation is available at http://mitarbeiter.fit.fraunhofer.de/~niemann/VisLA/

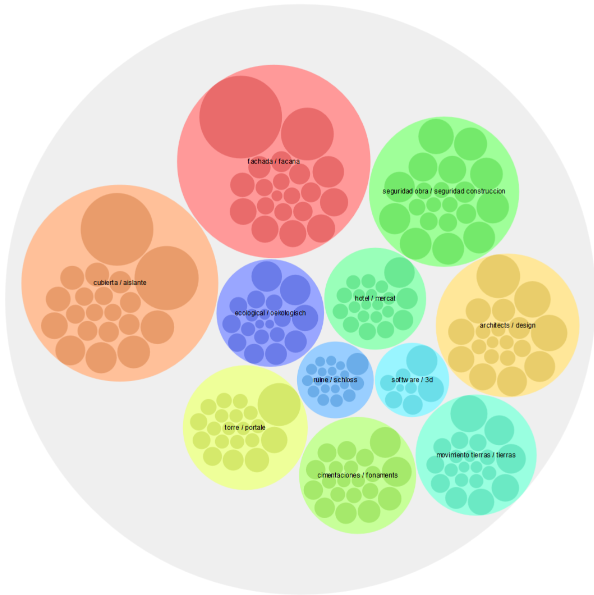

Figure 1: Start view

ods, objects, and materials used for insulation, e.g. *cobertes (covered)*, *paneles (panels)*, as well as *poliestireno (polystyrene)* and reference 683 distinct resources. The resources' descriptions hold further tags that can be used to orientate in this field. For example, figure 3 shows an excerpt from the list of resources that are assigned with the tag *sandwich*. While this tag might be unexpected at a first glance, the tags it was used with clarify its meaning, i.e. a (panel) structure made of three layers. Overall, 2,190 distinct tags are given in the resource list of this cluster.

Cluster 2: fachada / facana (facade). This cluster mainly holds Spanish and Catalan tags that deal with the construction and cladding of buildings, e.g. *sistemas constructivos (building systems)*, *cerramientos (enclosure)*, *gres (stoneware)*, and *constructivos (building)*. Overall, this cluster's tags reference 661 distinct resources that are assigned with 2,080 distinct tags.

Cluster 3: seguridad obra / seguridad construccion (work and construction safety). This cluster holds a mix of Spanish and English tags that deal with security, e.g. *seguridad trabajador (worker safety)*, *construction security*, *sistemas de seguridad (security systems)*, and *normativa (regulations)*. Overall, it references 532 distinct resources that hold 634 distinct tags.

Cluster 4: architects / design. The first cluster that mainly holds English tags and few Spanish ones deals with (green) architecture in the public space, e.g. *architecture, museum, green architecture, architettura (architecture), piazza*, and *bioarchitettura*. It references 296 distinct resources that hold 962 distinct tags.

Cluster 5: movimiento tierras / tierras (land movement). This cluster comprises Spanish tags that deal with the preparation of building zones, e.g. *excavaciones (diggings)*, *maquinaria (machinery)*, *calculo (calculation)*, and

Figure 2: Zoomed cluster view with a selected tag

excavadora (excavator). It references 278 distinct resources that comrpise 1,201 distinct tags.

Cluster 6: cimentaciones / fonaments (foundation). This cluster mainly holds in Spanish and Catalan tags that deal with the construction and anchoring of buildings, e.g. *muro (wall), building, terreno (ground), zapatas (shoes)*, and *anclajes (anchors)*. Overall, this cluster's tags reference 210 distinct learning resources that are assigned with 831 distinct tags.

Cluster 7: torre / portale (tower and portal). The main topic of this cluster is sustainability although its two most frequent tags do not imply it. Further tags are e.g. *bio edilizia (bio building), solar*, and *sostenibilidad (sustainability)*. However, it can be seen in the resource list that the learning resources that are tagged with *torre* or *portale* also deal with this topic, e.g. the insulation of towers. Thus, this cluster exhibits a topical relation to the first one but in contrast, in mainly contains Italian tags. Overall, the cluster references 201 distinct resources and its resource list comprises 661 distinct tags.

Cluster 8: ecological / oekologisch. This cluster also deals with sustainability but with a stronger focus on the generation and recovery rather than on the conservation of energy. Furthermore, it mainly comprises German tags, e.g. *photovoltaikanlage (photovoltaic power station), waermerueckgewinnung (heat recovery)*, and *waermepumpe (heat pump)*. The cluster references 200 distinct resources that are assigned with 825 distinct tags.

Cluster 9: hotel / mercat (hotel and market). This cluster holds tags that reference resources dealing with (aesthetic) buildings in the in public space like *puente (bridge), rascacielos (skyscraper), puerto (harbour)*, and *hotel arts* as well as famous architects of those buildings, e.g. *Santiago Calatrava Valls* and *Norman Robert Foster*. Overall, the clusters references distinct 86 learning resources that comprise 107 distinct tags.

Cluster 10: software / 3d. The only cluster that contains less than 20 tags deals with the design of buildings using the computer and comprises tags like *cad (computer-aided design), rhino3d (CAD Software), tutorial*, and *programming*. The cluster references 72 distinct resources that are assigned with 274 distinct tags.

Cluster 11: ruine / schloss (ruin and castle). This cluster references learning resources that describe or depict buildings built in the *mittelalter (middle ages)* or *hochmittelalter (high middle ages)* in German regions like *pfealzer wald (Palatinate Forest)* and *rhein-lahn-kreis (Rhine Lahn circle)*. Consequently, all tags are in German. Overall, they reference 51 distinct resources that hold 96 distinct tags.

Concluding, the clusters mostly contain tags that indeed belong to the same subject area, though, they are not completely separated. For example, several clusters deal with sustainability. However, their tags are in different languages and they have different focuses, e.g. the generation vs. the conservation of energy or public vs. private buildings. Furthermore, this shows that sustainability is an important field in architecture. The other clusters reference resources that describe different construction phases (design of buildings, preparation of building zones, as well as construction and cladding of buildings), security issues, and notable buildings as study objects.

In numbers, the tags that hold their own circles in the visualisation reference 2,849 distinct learning resources, i.e. a third of all tagged learning resources in the MACE data set.

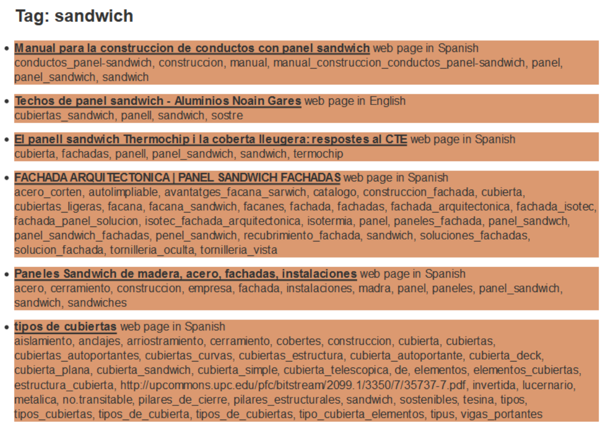

Figure 3: Excerpt of the resource list for the tag sandwich

While this number seems small at a first glance, it is quite high when considering that only about 3% of the tags hold their own circles. However, this number can be increased by presenting all resources referenced by a tag that was assigned to a cluster in the visualisation. So far, the tags that do not belong to the clusters' 20 most frequents ones are neglected.

Overall, the referenced learning resources are assigned with 6,585 distinct tags (i.e. half of all tags) which are shown in the resource lists. Considering that about 70% of the tags are only used once, this seems to be an acceptable number. Furthermore, it will be increased as well as soon as more resources are displayed.

7. CONCLUSION AND FUTURE WORK

In summary, the visualisation of the tag clusters gives a broad and easily understandable overview on the learning resources' subject areas. Furthermore, it enables the users to explore the data set by zooming into the clusters and browsing the result lists.

This visualisation, though, is not intended to be a standalone tool for the exploration of a data set. It is rather meant to be an additional tool that can be integrated with (already available) search functions like a faceted search or a social search as offered by the MACE portal. This way, the displayed resources could for example be filtered according to their language or media type and the tags in the resources' descriptions could be used to search for resources assigned with one ore more specific tags. Furthermore, the visualisation offers several possibilities for extensions. For example, by clicking on a learning resource in a tag's or a cluster's resource list, all tags that are assigned to this resource but are located in other clusters could be highlighted. This would further enhance the ability to discover relations between tags and, thus, between subject areas. Another option would be to allow the users to browse all tags belonging to one cluster and not only the most frequent ones.

So far, no evaluation has been conducted. In order to do so, the tag cluster visualisation needs to be integrated in a web portal. Thereafter, the acceptance of this visualisation can be evaluated by analysing its usage or by conducting a survey. Furthermore, user studies with control groups can be conducted to investigate if the use of the tag cluster visualisation increases the orientation in the portal or the performance of the students when solving tasks.

8. ACKNOWLEDGMENTS

The work presented in this paper has been supported by the Open Discovery Space that is funded by the European Commissio's CIP-ICT Policy Support Program (Project Number: 297229).

9. REFERENCES

[1] M. A. Hearst and D. Rosner. Tag clouds: Data analysis tool or social signaller? In *Proc. of the 41st Annual Hawaii International Conference on System Sciences*, HICSS '08, pages 160–, Washington, DC, USA, 2008. IEEE Computer Society.

[2] G. Heyer, U. Quasthof, and T. Wittig. *Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse.* W3L GmbH, 2006.

[3] O. Kaser and D. Lemire. Tag-cloud drawing: Algorithms for cloud visualization, 2007.

[4] S. Lohmann, S. Thalmann, A. Harrer, and R. Maier. Learner-Generated Annotation of Learning Resources - Lessons from Experiments on Tagging. In *Proc. of the International Conference on Knowledge Management (I-KNOW 2008)*, pages 304–312, 2008.

[5] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen. Getting our head in the clouds: Toward evaluation studies of tagclouds. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 995–998, New York, NY, USA, 2007. ACM.

[6] M. Scheffel, K. Niemann, S. Leon Rojas, H. Drachsler, and M. Specht. Spiral me to the core: Getting a visual grasp on text corpora through clusters and keywords. In K. Yacef and H. Drachsler, editors, *Proc. of the Workshops at the LAK 2014 Conference*, volume 1137 of *CEUR Proc.*, Indianapolis, Indiana, USA, 2014.

[7] S. Sen, J. Vig, and J. Riedl. Tagommenders. In *Proc. of the 18th international conference on World wide web (WWW '09)*, pages 671–680, New York, New York, USA, 2009. ACM Press.

[8] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. of the 17th international conference on World Wide Web - WWW '08*, pages 327–336, New York, New York, USA, 2008. ACM Press.

[9] J. Sinclair and M. Cardew-Hall. The folksonomy tag cloud: When is it useful? *J. Inf. Sci.*, 34(1):15–29, Feb. 2008.

[10] M. Stefaner, E. D. Vecchia, M. Condotta, M. Wolpers, M. Specht, S. Apelt, and E. Duval. MACE - Enriching Architectural Learning Objects for Experience Multiplication. In E. Duval, R. Klamma, and M. Wolpers, editors, *Proc. of the 2nd European Conference on Technology Enhanced Learning (EC-TEL '07)*, volume 4753 of LNCS, pages 322–336, Berlin, Heidelberg, 2007. Springer.

[11] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *In KDD Workshop on Text Mining*, 2000.