

The Hofmethode: Computing semantic similarities between e-learning products

Oliver Michel¹, Damian Läge¹

¹Arbeitsgruppe Angewandte Kognitionspsychologie, Universität Zürich

Key words: *e-learning, information retrieval, machine text understanding*

Abstract:

The key task in building useful e-learning repositories is to develop a system with an algorithm allowing users to retrieve information that corresponds to their specific requirements. To achieve this, products (or their verbal descriptions, i.e. presented in metadata) need to be compared and structured according to the results of this comparison. Such structuring is crucial insofar as there are many search results that correspond to the entered keyword. The Hofmethode is an algorithm (based on psychological considerations) to compute semantic similarities between texts and therefore offer a way to compare e-learning products. The computed similarity values are used to build semantic maps in which the products are visually arranged according to their similarities. The paper describes how the Hofmethode is implemented in the online database edulap, and how it contributes to help the user to explore the data in which he is interested.

1 How can one compare e-learning products according to similarity?

«Imagine that there is e-learning material out there, but nobody can find it!» A problem of the whole internet is reflected in the domain of e-learning business: The information is there, but we are unable to locate it. Although the situation is not yet so dire, the number of online repositories for e-learning material is increasing,¹ and consequently also the requirements of a suitable organizational structure. The coupling of repositories and standardization of an efficient set of metadata is one way to ease the path of information retrieval, but it is only half the battle. As soon as there is a large amount of data, this data needs to be organized if it is meant to be found by humans. Librarians, warehouse managers, and dictionary producers can tell us things. The same people can also tell us about the inherent difficulties. The information is normally organized by metadata, and experience has taught us that metadata are badly maintained. Keyword lists become old, new topics arise, different people use metadata in different ways, and categorization in general acts like the head of medusa. Not to mention the user, who might be completely unaware of the «appropriate» use of metadata search.

Therefore, the organizing structure should not (only) rely on abstract metadata, but should also involve the product itself, meaning the verbal full text descriptions or abstract of the product (which can be seen as part of the metadata too, but in contrast to other metadata, it is

¹ To name just a few of these: e-teaching (<http://www.e-teaching.org/didaktik/recherche/medienprojekte/index.html>), SWITCHollection (<http://www.switch.ch/de/els/collection/>), OCW (<http://ocw.mit.edu/OcwWeb/web/home/home/index.htm>), OpenLearn (<http://openlearn.open.ac.uk/>). These are not repositories themselves, but merely overviews!

the innermost core – the product has to be described somehow, unless we assume that the user already knows it).

If the organizational structure mainly involves the full text descriptions, then the key challenge is to compare the texts by similarity measures. What is similar? Similar compared to what?

The next difficulty affects the presentation of the data. Let us assume a user who is interested in an introduction to the different kinds of psychology. He enters some search text into a form of an e-learning repository. There might be several objects in the database which would suit his interest. How should the found objects be represented? An alphabetical list of titles is certainly far from being sufficient.

Semantic maps could form a promising approach: E-learning items are represented as dots on a map. Similar items are placed close together, and items that are further away from each other are dissimilar. This map metaphor seems to be cognitively well understood, as proven by various web services² and the authors' own experimental experiences. The user is able to explore the data intuitively (see Fig. 1) – the map provides help in finding what he is interested in.

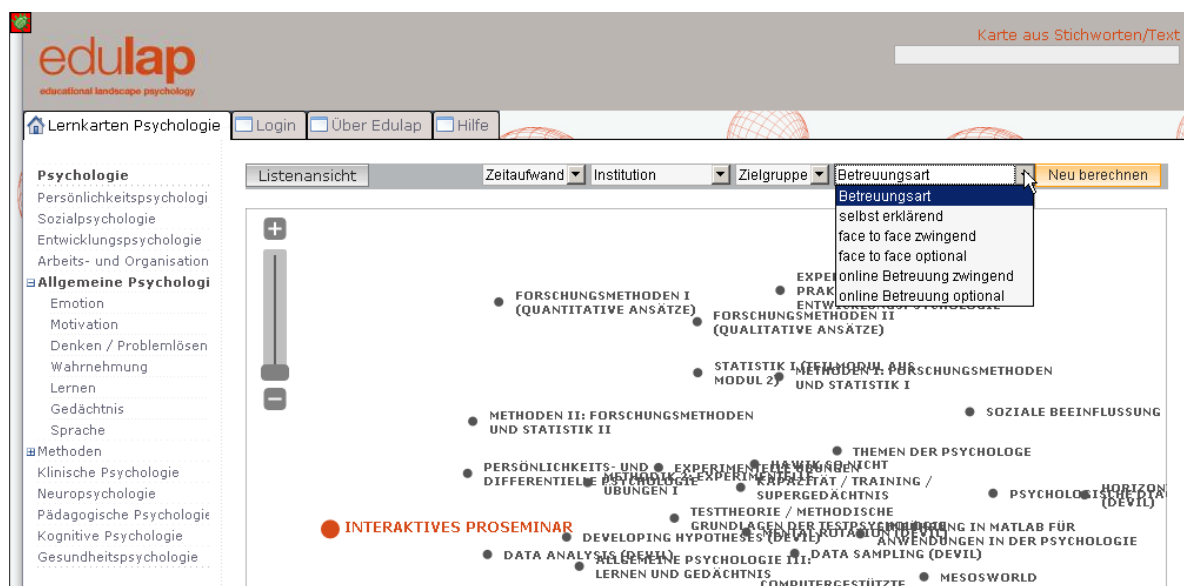


Fig. 1: Example of a graphical user interface, which is realized in the form of a semantic map. Close items are similar; items which are further away are dissimilar.

Again, the similarity measure is crucial. The key challenge is to find an algorithm which computes the semantic similarities between descriptions of e-learning products.

Two known methods for dealing with abstracts are *trigramming* and *Überlappungs-koeffizient* (ÜK) (overlapping coefficient). Trigramming handles text as an undistinguished stream of letters and divides it into packages of three letters. These trigrams are then compared to the trigrams of a second text. The more identical the trigrams, the more similar the texts.

The ÜK writes all different words of a given text into a table and compares the ratio of identical words with the table of a second text.

² Hulbee (<http://www.hulbee.de/>), Kartoo (<http://www.kartoo.com/>)

Both trigramming and ÜK share the disadvantage of being dependent on the specific style of the writer. Because both methods handle texts like bags full of isolated trigrams or words, they are unfocused, without any semantic concept. Trigramming reacts sensitively to whole formulations and phrases, which produce a high amount of identical trigrams. Thus, if a writer uses the same formulation in semantically different texts, trigramming is likely to overestimate the similarity. The ÜK, on the other hand, is highly dependent on the vocabulary used. Since every writer has his own style of writing and uses his own set of vocabulary, the ÜK is likely to distinguish writers rather than semantics. What is needed is an approach which focuses on semantically rich aspects of a text.

To solve the cited problems, we have developed an algorithm called *Hofmethode* (HM), which computes the semantic similarities of short to medium-sized texts. The resulting matrix of pairwise similarity values can then be used to generate a semantic map, in which the texts are ordered according to their semantic relationship.

The concept of the HM was proven to work in [1]. In this paper, we describe how the HM works and how it will be productively used in an online database of e-learning products.

2 How the Hofmethode estimates similarity between texts

The Hofmethode is an algorithm (based on psychological considerations) used to determine whether the meaning of a word in one text resembles the meaning of the same word in another text. Because the meaning of a word does not necessarily correspond to its shape, more information is needed. In fact, the word itself is almost useless if it is looked at as an isolated string. It is the context that gives the meaning to the word *in this specific situation*. Language utilization is fluid, not fixed. Therefore, the context of this word – referred to as *target word*³ – also has to be taken into account.

First we denoise the text by removing stop words and the like. Then, we extract the context of all target words in the text (some reflections on the compilation of the target word list are presented below). We define the context of a target word as the five words before and the five words after the target word. Because these words lie around the word like a halo, we call it the Hofmethode («Hof» is the German word for halo). These words are written into a table together with a value, which is dependent on their distance from the keyword: If it is the direct neighbour, the value is close to 1; if it is further away, the value lowers towards 0 (see Fig. 2), following a cosine function.

The context of the same target word in another text is also written into a table with the same procedure as described above. Now, the words in the two tables are compared: If there are identical (or even similar) words in the two tables, their multiplied (and summed up) values compose a similarity value between the two target words. If the value is high, then the meaning of the target word in these two contexts is regarded as similar.

³ We use the term *target word* to avoid confusion with *keyword*, which often denotes the keyword field in metadata descriptions.

structural processes correlates intelligence models structure attitude			based general estimated intelligence process emotional ratings		
Text A	Association	Value	Text B	Association	Value
	structural	0.71		based	0.71
	processes	0.87		general	0.87
	correlates	0.97		estimated	0.97
	intelligence	1		intelligence	1
	models	0.97		process	0.97
	structure	0.87		emotional	0.87
	attitude	0.71		ratings	0.71

similarity value of the target word
«intelligence» in text A and text B:
 $0.87 * 0.97 = 0.84$

Fig. 2: Simple example of the Hofmethode with a halo size of three words: In the two contexts of the target word «intelligence» are the same/similar words. Their values are multiplied. The resulting value represents the semantic similarity of the target word in these (and only these) two contexts.

This procedure is carried out with every target word in every text of a defined set of text items: We determine its context and compare it with the context of an identical (or similar) target word in another text. In the end, we have a *triangular matrix* of summed up similarity values. These values are transformed into Euclidian distances by means of *NMDS* (nonmetric multidimensional scaling) [2] and arranged in a semantic map. Similar texts will be positioned close together, thus building clusters, while dissimilar texts will be positioned further away. The wonderful aspect of NMDS is that even texts which do not share any similarity between themselves, but which share a covariance over other texts, can be positioned close to one another.

The compilation of the target words is not dealt with in this paper, so some general thoughts on this aspect suffice here: As mentioned above, our approach should focus on semantically relevant words. The more common a word is, the better we can compare its halos. On the other hand, if a word is too common, its semantic significance lowers. Therefore, we need common words with a wide variance of denotative meaning. To detect such kinds of words, we use statistical approaches. There is also the possibility to use (additionally) the common keyword field, which almost every set of metadata includes (it is not detrimental if the field is mostly empty, because it is still an enhancement of the statistical approach). Furthermore, we do not want a large list of target words, because this slows down the halo computing. Of course, if we have too few target words, some texts might have no target words at all and will therefore have minimum similarity.

3 The implementation of the Hofmethode in a real-world application

The HM is implemented in *edulap* (educational landscape psychology) [3], an online database for e-learning products (the screenshot in Fig. 1 is taken from the *edulap* prototype). The development of the database is in progress, so it is not yet publicly accessible. The system's first application is in the psychological environment. However, it will be adaptable in conceptual and technical terms to other academic fields.

Teachers and students are able to search the data by filters (e.g. format, date, or level) or by keywords. The resulting subset of texts is arranged in a semantic map, as described above. Additionally, users have the option to enter whole texts in order to find similar texts. This search input is regarded as a text item in its own right. Every item of the arranged subset is compared with the input by the HM and coloured by *distribution-based colouring* [4]. In the final map, one (or more) clusters will be coloured, telling the user in an intuitive and efficient way where to explore the data.

4 How to estimate the quality of the maps

Of course, it is crucial to be able to estimate the quality of the maps. For this reason, we developed an *expert's map*: We pseudo-randomly selected 70 learning products from the edulap database. Experts in psychology (mainly staff members of our institute of psychology) rated the similarity of the items by means of *parallel* and *hierarchical sorting*.

Using parallel sorting, the expert puts all the items into categories. He is free to choose the number and type of categories. This sorting type is very efficient (it took about 15 minutes to categorize the items), and immediately brings the main cluster structure to light. However, it does not allow a further analysis of the structure within a category. The maps based on the parallel sorting tend to show very tight clusters.

The hierarchical sorting type is much more time-consuming (approx. 50 minutes for the 70 items). The expert is asked to divide the set of the 70 items into two subsets. Again, the discriminating criterion is at his discretion, meaning that the number of items in the two subsets is likely to be quite different. In the next step, the items from the first subset are presented to the expert, and he is again asked to divide them into two subsets. Then, the second subset from the first step is presented to the expert and so on, until all groups contain three or less members. This method normally produces less strict groups and therefore allows some insights into the structure within a category. Several experts are needed for this sorting type, because only the comparison between the experts allows a conclusion to be drawn regarding the groupings.

The individual ratings of the parallel and hierarchical sorting of the experts resulted in maps, which were averaged to build the semantic expert's map (see Fig. 3).

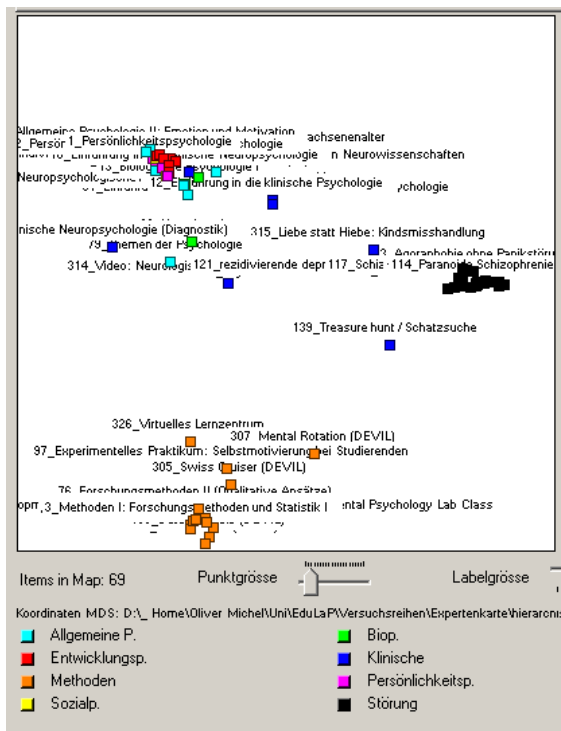


Fig. 3: The resulting expert's map: The colouring indicates the category of the item (orange: statistics and methods; black: mental disorders; other colours: various categories). Three tight clusters are apparent: The miscellaneous cluster in the upper left corner; the mental disorders cluster on the right; and the statistics and methods cluster in the lower left corner.

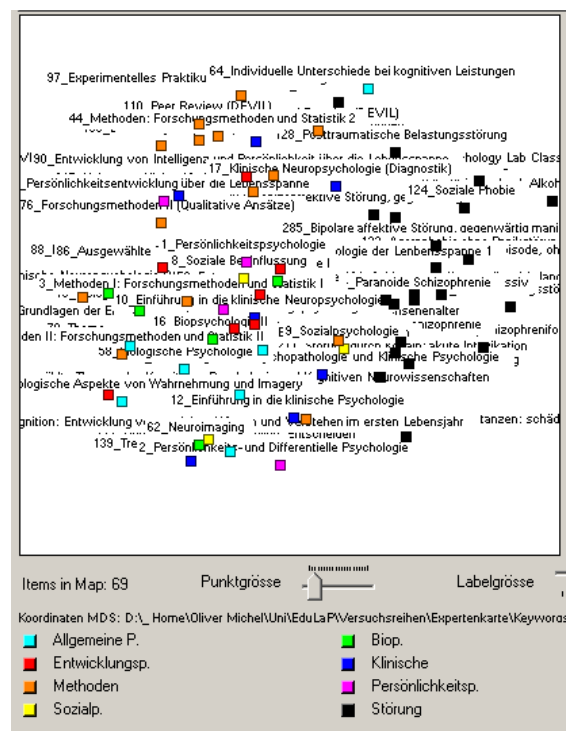


Fig. 4: The map based on similarity values of the Hofmethode. There are no clusters, but the map also reflects the ordering of the same three groups as the expert's map (the map has no orientation; it is the items in relation to each other that matter, not the absolute place on the map).

The set of the 70 items consisted of descriptions of mental disorders, a large part of statistical and methodological lectures, and some unspecified, remaining lectures. The sorting of the experts resulted in a map in which these three groups resulted in three tight clusters (the map of the parallel sorting produced even tighter clusters; the hierarchical sorting was less clustered, not shown in this paper).

Then, we took the descriptions of the same 70 items and compiled the target word lists using a statistical approach. The halos of all target words found in the 69 texts were computed and compared (one object had a very little abstract with no target words, so it was skipped). The resulting triangular similarity matrix was transformed into a semantic map by robust NMDS. This map shows a similar structure to the expert's group. The three groups (statistics and methods, mental disorders, various lectures) do not build tight clusters, but the items are nevertheless loosely grouped (see Fig. 4).

To make a more profound assertion about this resemblance, a *Procrustes transformation* is used, meaning that the two maps are laid onto each other, being rotated and resized until they achieve the best fit (see Fig. 5). It can be seen that about 6 items were in completely different places, although without necessarily being misplaced. The main difference between the two maps is the density of the main clusters: Whilst the high concordance of the experts leads to a rather categorical scheme (three dense main clusters), the similarity coefficients retrieved by the Hofmethode allows a more constant spread over the entire space, without losing the three general components. However, the two main groups on the left side start to overlap, which leads to a mixed section in the central left part of the map (see also figure 4).

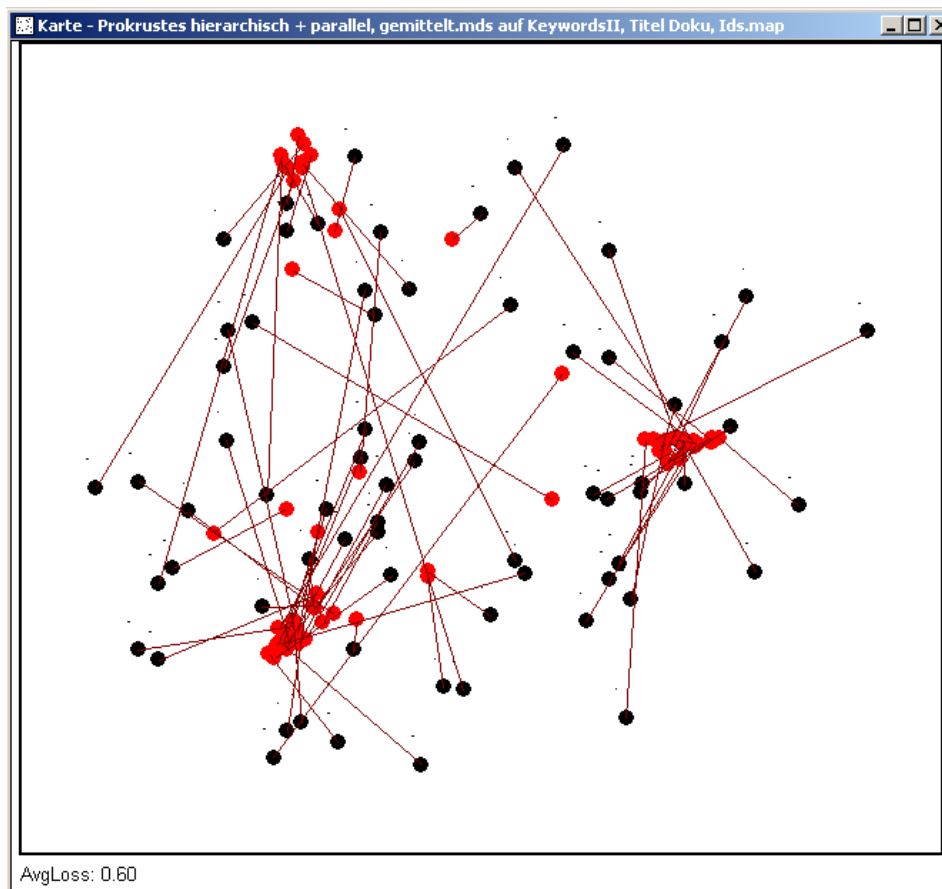


Fig. 5: The Procrustes transformation of the expert's map (red dots) and the Hofmethode map (black dots). The identical items from the two maps are connected with a brown line. The main structure is similar, although the expert's map shows tighter clusters.

The realistic scenario, which would lead to the HM map, would be a user filtering the edulap database for all products which are offered from a certain university. With one glance, he is able to obtain the whole field of products being offered. In a second step, he would use more filters, like format or learning time, or he would explore an interesting-looking part of the map and read the abstracts of the products. The map helps the user to find what he is looking for, by organizing the information in such a way that he is stimulated to explore it. The HM does not return him a list of ordered products, because what should be the ordering criterion? It should clearly not be alphabetical, and if it is to be the appearance of a certain keyword, then the question arises of which one, if the user hasn't yet entered one? Popularity? – Not a good idea, unless the user's only criterion was mainstream psychology. In a well-organized map, the user can grasp the whole field of available items. He can see the different topics and look at the products in the order he wishes.

This advantage could simultaneously be a disadvantage: The user does not get a solution in the form of a most preferred item. Instead, he *has to* explore the map himself. This demands some cognitive effort from the user and is not desired in every possible situation. In the environment of academic e-learning products, we expect this effort from the user, since alternatives demand even more power and time in the long term.

In short, the HM seems to organize the e-learning products automatically in a similar way to (averaged) experts. The HM is independent of most metadata, although it does need a full text description of the product.

Thus, the Hofmethode can be a fruitful help to automatically organize a sample of previously unorganized e-learning objects into a useful semantic structure.

References:

- [1] Michel, O. & Läge, D. (2006). *Die Hofmethode: Auf dem Weg zum maschinellen Textverständnis* (AKZ-Forschungsbericht No. 34). Zürich: Angewandte Kognitionspsychologie.
- [2] Borg, I., & Groenen, P. (1997). *Modern Multidimensional Scaling. Theory and Applications*. New York: Springer.
- [3] Streule, R., & Läge, D. (2008). Educational Landscapes: Mapping der elektronischen Ausbildungsangebote eines Faches mit Kognitiven Karten. In S. Zauchner, P. Baumgartner, E. Blaschitz & A. Weissenböck (Eds.), *Offener Bildungsraum Hochschule - Freiheiten und Notwendigkeiten* (pp. 50-57). Münster: Waxmann Verlag.
- [4] Ryf, S. & Läge, D. (2008). *Berechnung und Visualisierung von Verteilungen in NMDS-Karten am Beispiel des Musik- und Getränkemarktes*. In J. Reinecke & C. Tarnai (Eds.), *Klassifikationsanalysen in Theorie und Praxis*. Münster: Waxmann Verlag.

Authors:

Oliver Michel, lic. phil.
 Universität Zürich
 Psychologisches Institut
 Allgemeine Psychologie (Kognition)
 Arbeitsgruppe Angewandte
 Kognitionspsychologie
 Binzmühlestrasse 14 / 28
 CH - 8050 Zürich
o.michel@psychologie.uzh.ch

Damian Läge, Prof. Dr.
 Universität Zürich
 Psychologisches Institut
 Allgemeine Psychologie (Kognition)
 Arbeitsgruppe Angewandte
 Kognitionspsychologie
 Binzmühlestrasse 14 / 28
 CH - 8050 Zürich
d.laenge@psychologie.uzh.ch

