Latent semantic analysis as a tool for learner positioning in learning networks for lifelong learning

Jan van Bruggen, Peter Sloep, Peter van Rosmalen, Francis Brouns, Hubert Vogten, Rob Koper and Colin Tattersall

All authors are with the Educational Technology Expertise Centre (OTEC) of the Open University of the Netherlands. Jan van Bruggen is Educational Technologist and his current interests are computer-supported collaborative learning and application of latent semantic analysis in education (jan.vanbruggen@ou.nl). Peter Sloep is a Senior Educational Technologist. His interests are in technical affordances and social networks for distributed learning. Peter van Rosmalen researches how agents can support tutors in electronic learning environments. Francis Brouns and Hubert Vogten are Senior ICT Developers working in the areas of standards and the development of innovative e-learning environments. Rob Koper is Professor in Educational Technology. His research focuses on self-organized distributed learning networks. Colin Tattersall is Educational Technologist and his research interest are in innovation in e-learning and learning technology standardisation.

Abstract
As we move towards distributed, self-organized learning networks for lifelong learning to which multiple providers contribute content, there is a need to develop *new* techniques to determine where learners can be positioned in these networks. Positioning requires us to map characteristics of the learner onto characteristics of learning materials and curricula. Considering the nature of the network envisaged, maintaining data on these characteristics and ensuring their integrity are difficult tasks. In this article we review the usability of Latent Semantic Analysis (LSA) to generate a common semantic framework for characteristics of the learner, learning materials and curricula. Although LSA is a promising technique, we identify several research topics that must be addressed before it can be used for learner positioning.

Introduction
Educational practice is changing from being predominantly institution and teacher-led to being largely student-centred. We will not go into what this shift in emphasis entails, nor will we attempt to justify the need for it or to assess its current status. What we will discuss are the implications of such a change for a familiar problem: how can students be positioned with respect to the courses or programmes on offer? If we take the shift in emphasis seriously, then our conclusion must be that students – and not their educational institutions – will have to be able to assess their position. How can this be achieved?

To frame this question, we assume only that future learning environments will offer a diverse set of learning activities from which learners somehow may take their pick. The learner's history and goals define his entry position in relation to the learning activities. The learner's history, for example any learning activities that he has completed, will allow him to skip particular activities and to commit himself to completing others. A different entry position is likely to result in a different partition between activities to be skipped and activities to be completed. Different entry points will thus result in different paths through the set of activities being offered. The paths taken through the learning activities are likely to differ in their attractiveness to the learner, and in terms of the efficiency and effectiveness with which the learner achieves his personal learning goals. Some paths may also match a particular user's preferences better than other paths. Some users, for example, may want to study on their own - at their own pace, and where and when they choose - whereas others may want to study with a group of like-minded fellow students. Similarly, for a variety of reasons (practical, financial, pedagogical, and organizational) some paths may suit a particular provider better than others. A traditional, campus-based institution, for example, will have more trouble catering for the individual wishes of students than a distance teaching institution. In our definition, the positioning problem involves choosing an entry position that optimises efficiency, effectiveness and attractiveness, and reflects users' preferences and institutional constraints.

As an example of current practice, we will consider positioning at a traditional distance teaching institute, such as our home institution, the Open University of the Netherlands (OUNL). There are rules stipulating the admission criteria, and exemptions are granted on the basis of prior education and the transfer of credit points from obsolete OUNL courses or courses taken elsewhere. Even though only formally recognised prior education is taken into account, human expertise is still needed. Positioning here involves the direct mapping of student characteristics (goals, prior knowledge, history) onto the

characteristics of a curriculum and the individual courses within it, so as to identify the courses that the student can skip and those that the student still has to complete. The end result of the positioning process is that the student will be advised to follow a particular path through the remaining curriculum.

The OUNL still very much takes its own curriculum as the starting point. Its exemption policy really only serves to divide the set of courses that belong to a curriculum into two groups: those that the student in question needs to complete and those that he need not complete. The student's personal learning goals can only be met in so far as they can be mapped onto the set of available curricula. In other words, institutional constraints, rather than student preferences, determine which paths the student follows. It should be noted that precious few institutions differ from the OUNL in this respect. Positioning in a genuinely student-centred approach, however, cannot rely on the direct mapping of learner characteristics onto a curriculum. Curricula may exist, but cannot be the sole yardstick if the student is to have real freedom of choice with respect to their learning activities, the order in which they take place, and the pace at which the students engages in them. The institution nevertheless can and should let the student know his position in relation to the learning activities it has on offer; it should also advise the student on how best to proceed. This requires a knowledge of i) the learner, ii) the domain (the relevant learning activities), and iii) the order in which the learning activities are best carried out in view of the student's needs and preferences and the institutional constraints.

Human experts are capable of carrying out this positioning process, but even the limited positioning task undertaken by the OUNL already puts a considerable strain on available resources and staff time. Responding to positioning requests for each and every potential student and updating the information regularly rapidly becomes too heavy a burden. The only way out is to use an automated or semi-automated, computer-supported positioning process. This is no trivial task, however; in fact, it assumes that a considerable number of the research questions recently proposed for intelligent information systems have been answered (Cherniavsky and Soloway, 2002).

As if that were not complex enough, computer-supported techniques for positioning are feasible only if they meet at least two additional requirements. First, even when the shift towards student-centeredness implies that the outcome of the positioning process will be a recommendation, and not a formal admission decision, the institution still bears responsibility for making the recommendation. It therefore must be *reliable* (the same situation leads to the same recommendation) and *valid* (the recommendation matches that of experts). Secondly, the technique should scale easily, because it has to keep up with changing learning activities and growing numbers of learning activities, particularly if students are able to access learning activities from various providers. In the latter case, the learner's position with respect to a set of learning activities should cut across learning providers and learning domains. The problem is not merely one of increasing scale, then, but also of increasing complexity.

Anyone charged with the task of positioning learners in a learning network therefore faces a formidable set of hurdles, all of which need to be taken. This article addresses only two of them. We review how the inductive technique of Latent Semantic Analysis (LSA) can be used to generate domain and learner descriptions. We look into a number of LSA educational applications and describe the issues that, in our view, must be solved in order to use LSA in learner positioning.

Towards an LSA-based approach to positioning

LSA in a nutshell

Latent Semantic Analysis (Landauer *et al*, 1998) or LSA has its roots in research on document retrieval. By developing measures of semantic similarities between documents, LSA aims to improve retrieval beyond keyword matching (Dumais, 2003). The essential premise behind LSA is that co-occurrence of concepts in documents points to semantic similarities, ie, the documents, at least in part, address the same topics. We will not dwell on the mathematical details (see Deerwester *et al*, 1990, for a detailed description) but instead review the techniques involved.

LSA starts with a collection of documents. The frequencies with which the concepts occur in the documents are recorded in a table, the Concept-by-Document matrix. Note that the order of the concepts in the document is irrelevant: LSA does not log syntactic information. A document is represented by a column vector of concept frequencies (a document vector) and a concept is represented by a row vector of frequencies across documents (a concept vector). The dimensions of this Concept-by-Document matrix, let's call it X, are reduced in two stages. First, X is subjected to singular value decomposition (SVD). This is a process in which **X** is rewritten as the product of the matrices **C**.**S**.**D**. Think of **C** and **D** as defining orthogonal axes in a high-dimensional space and of **S** as a matrix that only has non-zero values on the diagonal. These values correspond to the length of the

axes. To reduce the number of dimensions, only the longest axes are retained by removing rows and columns in S and the corresponding ones in C and D. The original matrix is now reconstructed from these reduced matrices. In the *reconstructed* matrix, a document-vector may contain a frequency for a word W that did not appear in the original document. In other words, a query for "all documents about W" may return documents that do not contain the word W itself, but words that tend to co-occur with W. Several other measures can be obtained using the reconstructed Word-by-Document matrix, such as the correlation between document vectors. The higher the correlation, the more the documents resemble one another. That makes it possible to compare documents to each other, or compare a document to a vector of search terms. The LSA rating of text similarity closely matches those of human raters (Dumais, 2003; Wiemer-Hastings and Graesser, 2000).

From its original inception in information retrieval, LSA has found wide application in such research areas as cognitive models of human word meaning acquisition (Landauer and Dumais, 1997) and language understanding (Kintsch, 1998; Wiemer-Hastings and Zipitria, 1999). Our intention here is to review applications of LSA in educational settings. There are two main types of LSA educational applications. The first concerns assessing essays and providing feedback to students. The aims of the systems developed here may vary from providing a summative evaluation to offering formative support to students who are preparing essays or summaries. The second type of LSA educational application involves modelling the knowledge of the learner in order to select and sequence suitable instructional materials. Here, LSA is used to model both learners and instructional materials in the same multi-dimensional semantic space, making it possible to assess similarities between the two. Selecting material that is in the "zone of proximal development" is the key challenge in this type of application. We shall discuss both kinds of application in turn.

Support and evaluation of written composition
The LSA application that is probably best known in the educational community is the Intelligent Essay Assessor (IEA), a program that rates student essays (Foltz *et al*, 1999). As with any LSA application, the Intelligent Essay Assessor is trained on material drawn from the domain of the essay topic. IEA does not require a large set of graded essays. Tuning the system may require just a few examples, including a so-called "golden standard". IEA has been found to rate essays with a reliability that matches those of human raters.

To solve the positioning problem, we are not interested in essay rating as such but in the mechanisms by which LSA identifies document characteristics and compares documents to a domain model. LSA can be used to measure several aspects of texts. Coherence is measured by calculating the similarities between the individual sentences. A high overall similarity indicates repetition or rephrasing of the text, while an overall low similarity is an indication that the text has a low coherence. Drops in similarities between successive sentences indicate topic breaks and a high average number of topic breaks indicates that the text jumps from topic to topic.

These types of measures are used in Select-a-Kibitzer (Wiemer-Hastings and Graesser, 2000), an interactive tool that provides feedback on student compositions. Once a student has entered the text, specialist agents - the Kibitzers - may be invoked to provide feedback on the text characteristics in which they specialise. LSA is used to determine the coherence of the text, and the topic breaks between the sentences are used to identify semantic chunks. The sentence with the highest average similarity to other sentences in the chunk is considered the key sentence and is presented as the system's understanding of the topic. The LSA engine in Select-a-Kibitzer is trained in template sentences to help determine the purpose of sentences. Templates such as "I would change…. because" are used to indicate why-reasoning.

The techniques discussed here allow us to represent the domain, the learner, and the learning materials in a single semantic space. The semantic space of the domain is created using several texts, including, but not limited to, the texts that the learner may have to study. Learners may be represented, as in IAE, by one or more of the texts they have produced or by the similarity of their text to standard texts. They may also be represented by a set of vectors of the texts they studied. Finally, they may also be modelled using proxies, for instance descriptions of the learning activities they have completed, rather than the activities themselves. The learner vector or vectors can be compared to the vector or vectors of a learning activity to determine how they correlate. A high correlation indicates that the learner can skip the activity. Learning activities can also be compared to one another and a large degree of similarity indicates considerable overlap in the subject matter.

Some caveats are in order, however. All LSA models are based on co-occurrence of concepts in documents. The *order* in which these concepts occur/co-occur is completely ignored. As Wolfe and Goldman (2003) point out, LSA fails to represent domains in which the context determines how sentences should be interpreted. This applies to domains that use metaphorical language, causal reasoning and logically ordered sequences of steps. It remains to be seen whether using LSA in an educational context with fixed sequences of learning activities will mitigate this problem. Some of the research discussed below indicates that this may indeed be the case.

Learner modelling and selection of instructional material
As indicated, the second way in which LSA is applied in an educational setting involves the selection and sequencing of instructional materials. AutoTutor (Graesser *et al*, 2000) engages students in a natural language conversation and thus encourages them to provide elaborate answers to the questions it poses. AutoTutor scores the quality of the answers that the students provide in conversational turns using a variety of techniques, including LSA. AutoTutor rates the quality of the students' assertions much the same as intermediate-level experts, but not as well as accomplished experts. The LSA component of AutoTutor is able to discriminate between classes of simulated students, and is capable of tracking the increased coverage of a topic in successive turns.

Zampa and Lemaire (2002) used LSA in an intelligent tutoring system to model the domain and the student. In their model, a domain is built of "lexemes", being either words in a language-learning domain, or facts and conclusions in a problem-solving domain. Note that this domain representation is not based on raw text, but requires prior identification of the lexemes. The student, it is assumed, learns the domain by being exposed to a series of lexemes. The tutoring system selects those texts/topics in a zone around the student and domain sequences that have already been addressed. Sequences that are too close or too remote are expected to yield a weaker learning effect and are therefore ignored.

Wolfe *et al* (1998) addressed this same issue, referring to it as the "zone-of-learnability". The key to their approach was a study text selected to match the prior knowledge of the learner as closely as possible. First, they collected data on the students' prior knowledge, and then had the students study one of four different texts about, say, the anatomy, function and purpose of the human heart and the circulatory system. The texts ranged in difficulty from elementary school to medical school level. As expected, learning gains were related to prior knowledge: texts that were too easy or too complex yielded weaker learning gains. Wolfe *et al* (1998) presented a number of curve-fitting solutions that relate LSA-based similarity measures between prior knowledge and the study texts to predict learning effects.

A prototypical tool (HEADHUNTER) that matches jobs, people and instruction is worth mentioning here. Laham, Bennett and Landauer (2000) processed data on three Air Force occupations for which full job descriptions were available. They then analysed these "duty lists", tasks grouped into functional units, and individual tasks along with the tasks actually completed, thereby constructing a single semantic space for jobs and people. The semantic similarities between jobs and people could be used to decide between candidates for the job or to select a replacement. Laham *et al* (2000) indicate that, adopting the approach used by Wolfe *et al* (1998) discussed above, the system could also be trained to predict which course would bring a person closer to a target job profile.

In view of these experiences, LSA techniques would appear to be useful for learner modelling as well as for the selection and sequencing of instructional material, thus addressing important aspects of the positioning problem sketched above. A cautionary note is nevertheless in order. First of all, the granularity of most of the studies presented is finer than needed for positioning. Most of the applications, possibly with the exception of HEADHUNTER, used relatively small domains. The researchers sometimes pre-processed the texts used to train the LSA to mould the semantic space in such as way that it could support comparison of learners' texts with templates or lexemes. Consequently, LSA could be used to track progress in AutoTutor or model a zone of proximal development.

Positioning does not require us to model the student or the learning activities in such fine detail that we can track student progress. A level of description suffices that is *detailed* enough to differentiate between different learning activities and that is *broad* enough to cover the contents of a curriculum (at the level and extent of, say, a bachelor's programme). This approach may result in too many documents in the domain and in the learner's history correlating. The risk is that learning activities will no longer be discernible. However, Foltz *et al* (1998), who used LSA to measure coherence in

textbooks, found that the coherence within two introductory psychology textbooks remained at a relative stable level throughout the pages of the volume. This, we surmise, indicates that we may be able to differentiate between learning activities if their contents can be used to train LSA.

Using Latent Semantic Analysis for learner positioning

We have discussed a number of LSA educational applications and noted that these applications used models and measures that hold promise for the problem of learner positioning. No current LSA application specifically addresses the positioning problem, however, and we must therefore consider a number of issues that must be addressed before LSA can support learner positioning. Our review will reveal a number of limitations of LSA, discussed in the same order as the previous sections.

Creation of a single semantic space

LSA creates a semantic space of a domain in which learning activities and students can be modelled and compared. In LSA, a domain is modelled by deriving a high-dimensional semantic space from materials, such as textbooks and articles that cover the domain. A student is modelled by mapping the student's documents (history) onto the semantic space. The similarities between learning activities are calculated as the correlation between their document vectors. Similarly, high correlations between learner vectors and activity vectors indicate that the learner has already mastered the content. This would identify activities that the student can skip.

The first issue to consider is whether LSA is applicable to all domains. We have already mentioned the inability of LSA to deal with context dependency. In domains with high levels of context-dependency, the use of LSA may be limited to and analysis of *descriptions* of the contents of learning activities and other materials. This limitation will also apply for the construction of learner models, in particular the processing of additional learner data.

A second issue relates to the specificity of the learning materials used for LSA. There are clear indications that LSA works better if the semantic space contains *specific* materials. This seems obvious where LSA is used to guide learning (as in AutoTutor). For the positioning problem, the trade-off is more complex: we need general level material to represent the curriculum and the domain as a whole. We may need general descriptions of the learning material to position it in the domain. On the other hand, we need detailed, specific materials to represent the content of the learning activities in order to map the student's prior knowledge onto the material of other learning activities.

A third issue concerns the stability of LSA parameters. As described above, we assume that student-centred learning environments are highly dynamic: they will frequently change and become more complex. If LSA is to help solve the positioning problem in student-centred learning, it must require only minimal support when material is added, altered or deleted and the environment changes as a result. The computational load of LSA is considerable, but that is not the biggest issue here. More serious is the question of the stability of parameters. Only by conducting research can we establish stable parameters for the optimum number of dimensions of the semantic space, the vector cosines to be used, and the predicted knowledge growth when additional materials are added to the semantic space.

A fourth issue is related to the openness of student-centred learning. Openness matters, as it enables a variety of different parties, including the students, to contribute to learning activities. This implies, however, that competition between learning activities may arise. Although the research on LSA discussed above indicates that coherence measures can be used to identify topical areas in the texts, it remains to be seen whether these can be used with other LSA-based measures to compare the contents of competing activities.

Openness also implies that activities may have different source languages. Although LSA has been shown to deal effectively with translations between vocabularies, there is, to our knowledge, no proof that LSA can be used to map domain descriptions in different languages. LSA might be able to deal with this issue in domains such as mathematics, science, and psychology, which use symbols and formulae and are therefore less dependent on natural language, but it is likely to run into problems in domains in which this is not the case.

Identification and sequencing of learning topics.

How well can LSA order the activities that the student should complete? Although examples were discussed in which LSA sequenced learning materials, it remains to be seen whether they can be generalized to the sequencing of complete learning activities. In the examples, LSA operated in small

domains and the researchers prepared curricular sequences or identified topics such as the "lexemes". This is certainly not a sustainable approach in the open environment we envisage. Neither can we assume that we will find explicit representations of learning activities and learning trajectories.

Nevertheless, several approaches provide interesting material for further research. One possibility is that providers can use metadata descriptions to indicate the degree of difficulty for materials or to describe supposed prior knowledge in terms of courses. The IEEE learning object metadata specification does provide slots for such a description (IEEE, 2002). If so, one of the topics of future research might be to turn these statements into vectors within the semantic space in order to compare them to vectors representing the knowledge of the learner or the contents of other activities.

Another approach, admittedly a rather speculative one, might be to implement the curve-fitting methods discussed above for complete activities and predict learning outcomes on the basis of similarities between prior knowledge and the contents of learning materials or their description. Additional work is needed to determine whether, in the long run, it will be possible to gather sufficient data to produce an accurate prediction.

A final issue for research, obviously, is to establish how valid the LSA-based positioning recommendations are. LSA bases its positioning solutions solely on similarities and expected knowledge growth. The examples of predicted growth that we have reviewed were all based on comparisons of topics within a small domain. In the case of positioning we are, potentially, comparing learning activities of different qualities. LSA may have to weight quality criteria or the system may need human expertise to guide the learner. Note that the approach to positioning described here can also be used to provide important inputs into current exemption procedures, because it can be used to evaluate student inputs (essays) or material previously completed by the students in terms of the semantic space of the curriculum and its learning material.

Conclusions

We have reviewed Latent Semantic Analysis (LSA) as a technique that may help us solve some of the positioning problems we are bound to encounter once we implement student-centred learning. LSA offers a drastic approach by computing a completely abstract semantic model of the domain, the contents of the activities and the learners. It offers us models of the learner, the domain and the curriculum, but these cannot be inspected or modified in the traditional sense: the model consists of vectors, similarities between vectors and measures derived from them. The abstract nature of the semantics makes it possible to avoid maintaining extensive descriptions and mappings that would otherwise be needed. Our review of LSA indicates that it may indeed help solve the positioning problem. However, we also identified several areas in which further research is required before LSA can begin to deliver on this promise.

Acknowledgements

The authors wish to thank the management and staff of the Schloss Dagstuhl International Conference and Research Centre for Computer Science for providing a pleasant, stimulating and well-organized environment in which to write this article.

References

Cherniavsky J C and Soloway E (2002) A survey of research questions for intelligent information systems in education *Journal of Intelligent Information Systems* 18, 5-14.

Deerwester S, Dumais S T, Furnas G W, Landauer T and Harshman R (1990) Indexing by latent semantic analysis *Journal of the American Society for Information Science* 41, 391-407.

Dumais S (2003) Data-driven approaches to information access *Cognitive Science* 27 3, 491-524.

Foltz P W, Kintsch W and Landauer T K (1998) The measurement of textual coherence with latent semantic analysis *Discourse Processes* 25 2-3, 285-307.

Foltz P W, Laham D and Landauer T (1999) Automated essay scoring: applications to educational technology *Edmedia 99*. Retrieved online 02/12/02 at: http://www-psych.nmsu.edu/pfoltz/reprints/Edmedia99.html

Graesser A C, Wiemer-Hastings P, Wiemer-Hastings K, Harter D and Tutoring Research Group (2000) Using latent semantic analysis to evaluate the contributions of students in autotutor *Interactive Learning Environments* 8, 129-147.

IEEE (2002) *Draft standard for learning object metadata*. New York, NY: IEEE. Retrieved online 30/03/2004 at: http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

Kintsch W (1998) The representation of knowledge in minds and machines *International Journal of Psychology* 33, 411-420.

Laham D, Bennett W and Landauer T K (2000) An LSA-based software tool for matching jobs, people and instruction *Interactive Learning Environments* 8, 171-185.
Landauer T and Dumais S T (1997) A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge *Psychological Review* 104, 211-240.
Landauer T, Foltz P W and Laham D (1998) An introduction to latent semantic analysis *Discourse Processes* 25, 259-284.
Wiemer-Hastings P and Graesser A C (2000) Select-a-Kibitzer: a computer tool that gives meaningful feedback on student compositions *Interactive Learning Environments* 8 2, 149-169.
Wiemer-Hastings P and Zipitria I (1999) Rules for syntax, vectors for semantic in Moore J D and Stenning K (eds) *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society* Lawrence Erlbaum Associates, Mahwah, 1112-1117.
Wolfe M B W and Goldman S R (2003) Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions *Behavior Research Methods Instruments & Computers* 35 1, 22-31.
Wolfe M B W, Schreiner M E, Rehder B, Laham D, Foltz P W, Kintsch W, Landauer T K (1998) Learning from text: Matching readers and texts by latent semantic analysis *Discourse Processes* 25 2-3, 309-336.
Zampa V and Lemaire B (2002) Latent semantic analysis for user modeling *Journal of Intelligent Information Systems* 18, 5-14.