Appeared in 2004, *Cognitive Systems*, *6*(2-3), 227-237.

SIMULATING STUDENT COMPREHENSION WITH LATENT SEMANTIC ANALYSIS TO DELIVER COURSE READINGS FROM THE WEB

Philippe Dessus

Laboratoire des sciences de l'éducation & IUFM (Institut Universitaire de Formation des Maîtres), Grenoble, France

Abstract

Providing students with appropriate additional texts from a course (i.e., readings) requires cognitively demanding human activity. Moreover, the content of such readings might not be adequate to the ongoing student understanding of the course. Systems that provide such readings are either traditional hypertexts using links designed by hand, or adaptive hypertexts using a sophisticated model of the student. The method described here uses LSA, a statistical procedure devoted to the semantic comparison of texts. The method mimics some characteristics of human discourse understanding through a simplified version of the Construction-Integration model. This method thus retrieves course readings from the web that take into account the ongoing comprehension of the course by the students.

Abbreviations: LSA latent semantic analysis; AH adaptive hypertexts

1. Introduction

Providing adequately students with "the next text to be read" is a well-known problem in the field of Intelligent Tutoring Systems. Various strategies have been implemented to tackle this problem, focusing alternately on domain content, student prior knowledge, or pedagogical considerations. We focus here on domain content by proposing a method that automatically delivers readings to students (i.e., additional texts close to the course content). The automatic delivery of such documents is a twofold problem: 1°) in order to be retrieved, the entire corpus either has to be coded beforehand, or has to be subject to an information-retrieval method; 2°) in the latter case, another problem arises: university courses are usually short and very specialized thus they would be not convenient because information retrieval methods usually require large corpora. Any large corpus (e.g., encyclopedia) would be not convenient, being related to a general content. Then, an appropriate way to retrieve course readings would be to perform web queries.

Revised version of a poster presented at the Special Workshop on Multidisciplinary Aspects of Learning of the European Society for the Study of Cognitive Systems, Clichy, 17-19 January, 2002.

We present here a system in which "the next text to be read" can be extracted from Internet, through a model of ongoing student comprehension of the previous texts read. This model uses LSA (Latent Semantic Analysis), a statistical procedure devoted to the semantic comparison of texts (Landauer & Dumais, 1997). This way, since the delivered texts are compared with the database texts on-line, the prior typing of the delivered texts is not necessary. Such a system could help teachers provide students with readings that are closely related to their comprehension of the course content. The texts of the readings being related both to the students' ongoing understanding of the course and the course content, through appropriate hypertext links. The remainder of this paper is as follows: section 2 reviews some trends about hypertext links generation, section 3 presents the Latent Semantic Analysis, which is the core of our method. Some studies using LSA to model knowledge are described in section 4. Finally, sections 5 to 8 detail our method.

2. **From Hypertext Links Generation to Adaptive Hypertexts**

Our method pertains to two research fields: automatic hypertext links generation and Adaptive Hypertexts. The first field is devoted to the generation of hypertext links according to the document structure or content. The second field takes into account students characteristics, either static (i.e., experience, knowledge, attitudes, and so on) or dynamic (i.e., previous pages read, number of activated links, and so on) to dynamically provide pages that are more adequate. The main features of these fields are as follows.

Automatic hypertext link generation is commonly based on document surface features. Content tables, references, indices from simple structured documents are used to create structured HTML documents (Quint & Vatton, 1992). More sophisticated methods take into account deeper text features to generate links. For example, information retrieval methods are used to represent each word of the document by a vector (each document being represented by the sum of the vectors of each word). In this way, it is straightforward to compose a query by comparing its vector to all the document vectors. The closer the two vectors are, the stronger the corresponding documents are semantically related (Salton & McGill, 1983). This kind of method has been positively tested to generate hypertext links (Blustein & Webber, 1995; Goffinet & Noirhomme-Fraiture, 1996; Landauer et al., 1993), even for non-structured documents containing subtle cross-referencing clues (i.e., see *X*; cf. *Y*, and so on). It is worth noting that these methods use either document structure or document content to generate HTML code, they thus do not take into account any student model.

From this first research field, mainly concerned with developing tools for document creation, derived a second one, more concerned with helping user navigation. Adaptive Hypertext (henceforth AH), aims at integrating some aspects of student knowledge into HTML production (Brusilovsky, 1996; de Bra, Brusilovsky, & Houben, 1999), in order to provide users with an interface that is functionally and cognitively adequate to their navigation activity. Many applications in the field of educational hypermedia have been developed by using this approach. They share some features we detail now (see de Bra

et al., 1999). The formalization of knowledge contained in a AH is twofold: *concepts* are connected through *relationships*. AH tracks all user actions (i.e., browse pages, activate hypertext links) in order to maintain a model of each user. This model is used to rank all pages according to the user's current knowledge or interest. Two kinds of content selection are then performed. Firstly, the links to the most relevant pages are then dynamically enabled; the less relevant pages from the user point of view are disabled conversely. Secondly, the content displayed into pages can also be enabled or disabled according to the user's profile.

While traditional hypertext systems deliver the same document whatever the student profile is, these systems generate personalized documents. Some drawbacks remain however:

– In AH, the knowledge formalism often used is knowledge map. The assumption that such maps mimic human knowledge is also often supported. We claim that is a poor way to represent knowledge: —knowledge cannot be represented by a two-dimensions map without considerable loss; —the connections between words or concepts are emphasized without appropriate account for words and concepts themselves.
– Relations between nodes (pages) and links are often represented by semantic networks. Such relations require to be hand coded, therefore they are time-consuming and domain-dependent.
– A dynamic interface may also be viewed as problematic: hiding, disabling or removing both content and links do not respect interface consistency, which is considered an important prescription in Computer-Human Interaction research (Payne, 1991).

Our method differs from the two approaches described above. Firstly, this method mimics some aspects of human comprehension of texts read. Secondly, this method uses a high-dimensional space to represent knowledge. Thirdly, this method neither generates static HTML documents nor extracts documents from a textbase according to specific student profiles. This method rather dynamically generates hypertext links to further "readings" related to the ongoing course content (i.e., the amount of texts already read). The semantic matching between course content and readings is computed by Latent Semantic Analysis, a method we describe now.

3. **Presentation of Latent Semantic Analysis**

Latent Semantic Analysis is a factorial analysis method based on word co-occurrences that relies on large amounts of texts to build a high-dimensional semantic space. We will not detail here the mathematical aspects of LSA which are presented elsewhere (Deerwester et al., 1990; for comprehensive presentations of LSA see Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). An intuitive viewpoint of how LSA works is formulated as follows (Lemaire & Dessus, 2003):

(1) Two contexts (i.e., sentences, paragraphs, or texts) are semantically similar if they contain similar words.
(2) Two words are semantically similar if they appear into similar contexts.

This mutual recursion is solved by a singular value decomposition—a statistical method close to a factorial analysis—to reduce the huge word-by-document matrix to about 300 dimensions. The conceptual closeness between each word or paragraph of the corpus is computed as the cosine between two vectors representing words or paragraphs in the reduced space. For instance, *assessment* and *evaluation* are semantically similar since they appear in similar contexts (i.e., paragraphs). They occur with similar words like *term*, *papers*, *grades*, *teachers*, and so on. The way LSA processes texts—each paragraph is considered a "bag of words"—has its own strengths and weaknesses (Rosé et al., 2002). LSA requires a limited human pre-processing (i.e., corpus coding, software development), it is insensible to ungrammatical input, its performance is relatively close to human's one. The LSA's main weakness is to be less precise than linguistic knowledge based approaches, and to capture less aspects of language (i.e., syntax, negations, rhetoric or style). However, various tests in the domain of language show that LSA's performance is close to human subjects' performance (Landauer et al., 1998). We present now some tests that focus on the modelling of human comprehension.

4. **Modeling Knowledge from Text with LSA**

Many experiments have shown that LSA is an adequate model of knowledge acquisition through exposition to texts.

Wolfe et al. (1998) used LSA to simulate student knowledge acquisition. Firstly, LSA built a semantic space using texts from a handbook on the anatomy and the physiology of the human heart. Secondly, students were asked to write texts about the human heart, in order to measure their prior knowledge. These texts were added to the previous semantic space to be compared with the texts from the handbook. The students' texts also were graded by human judges. Thirdly, students were asked to answer [leave out this 'to'] a general questionnaire about the heart. Good correlations (correlation coefficients between $r = .63$ and $r = .74$) were found between human grades and similarities computed by LSA; and between the latter similarities and the general questionnaire grades.

We replicated (Dessus, 2000) the experiment of Wolfe and colleagues with a more dynamic text input, in french. We simulated the ongoing student comprehension while they were taking a course. The answers of the students to a MCQ test were correlated to LSA's predictions. We obtained a smaller correlation ($r = .30$) compared to the original data. This result may be explained by the lower amount of text—MCQ vs. handbook—processed to perform the matching.

As impressive as they may seem, these results are limited to the number of texts processed by LSA. To our knowledge, the largest corpus ever analyzed by LSA represents about 24 millions words (149 megabytes). Since such a corpus takes many

hours to be computed, we may consider another way to retrieve text that is related to a content. The web could be such an appropriate source of texts, as Turney (2001) has shown.

Turney performed a comparison between LSA's performance and a set of web queries. He showed that a simple algorithm using a search engine obtained a better score than LSA at a 80-items TOEFL test (about 10% higher). This performance can be attributed to the huge amount of texts the search engine uses to retrieve data. This result shows that although LSA uses a considerably smaller amount of data, its performance compares with the performance of web queries. These experiments show that LSA is an appropriate tool for simulating some aspects of human comprehension. We describe now more precisely the process of that simulation.

5. **Simulating Student Course Comprehension with LSA**

The purpose of our method is to automatically deliver texts to students during their reading of a course. It is straightforward to complete such a delivering by hand, by composing a web query with the titles of the different parts of the course. In that case, the relevance of the query is too much dependent on the formulation of the titles. Our method both takes into account the entire course content (i.e., not only its titles), and the individual comprehension of the course by the students, to retrieve readings from the web.

Comprehension can be viewed as a cumulative and iterative process. Kintsch (1988) formalized the process of discourse comprehension in two iterative steps: *Construction* of a textbase of inter-related units (i.e., propositions) and *integration* of the most relevant propositions by resolving ambiguities. More recently, Kintsch (2001) used LSA to propose a simulation of the Construction-Integration model using LSA. We describe here a simplified version of this model:

— *Construction of a textbase*. Consider $\mathbf{E}_n$, a course excerpt (i.e., the n first paragraphs of the course). Let $\{\mathbf{S}\}$ a set of paragraphs representing the domain knowledge (i.e., educational research conference abstracts), each paragraph being an abstract. These latter paragraphs can be sorted in terms of their semantic neighborhood with $\mathbf{E}$ (i.e., their relatedness to $\mathbf{E}$). One can select a subset $\{\mathbf{s}_n\}$ of $\{\mathbf{S}\}$, which can be viewed as an associative network of related ideas developed in $\mathbf{E}_n$. This subset is called "short readings".

— *Integration of knowledge*. $\mathbf{a}_n$, the closest abstract to $\mathbf{E}_n$ in $\{\mathbf{S}\}$ is selected to extract its title. At each moment of the course, $\mathbf{a}_n$ can be considered an approximation of the current knowledge activated during the course reading.

— *Large readings search*. Thus, at each moment of the course, the title of the current $\mathbf{a}_n$ can be extracted and be used to formulate a query that retrieves documents from the web. This new set of retrieved documents is called "large readings". Due to the huge amount of documents retrieved with such a query, the two most recent abstracts were used to perform the query (i.e., $\mathbf{a}_{n-1}$ and $\mathbf{a}_n$).

We implemented this method in the Perl programming language running on a Unix station according to the following steps:

1. A course text (on educational sociology, 0.1 megabytes) and the textbase (about 1,900 abstracts of an educational sciences conference, 3.8 megabytes) are computed in the same 300-dimensions semantic space. Each paragraph of the text course is then represented by a 300-dimensions vector, allowing semantic comparisons with textbase documents.
2. A simulation of the course is performed, by adding successively a new course paragraph to the previous ones, and by retrieving at each addition the closest textbase document to the ongoing course. All these textbase documents constitute the short readings base.
3. The titles of each document retrieved from the textbase in *Step 2* are extracted.
4. The document titles allow to retrieve Internet documents (i.e., large readings base). By adding successively a new word to the query, one can retrieve documents close to the course content through an Internet query.

6. **Results**

We performed this method by using a course on the sociology of education. Results are presented in Table I and Figure 1 below. Two documents are sufficient to retrieve about a set of about a hundred large readings related to the course. The student can read some documents from this set (see Table II the list of the first ten documents). However, this measure is only quantitative: the content value of the retrieved documents remains to be rated by human judges.

Insert Figure 1 about here

Additionally, this method should model some aspects of the representation of the ongoing content of a course, from the viewpoint of the students. Figure 1 shows that, once sufficient amount of text has been processed (i.e., thirty paragraphs, representing half of the course), only two different documents are activated (documents #254 and #311). More generally, the total number of documents activated can be viewed as a measure of content diversity delivered by the course. The more diverse the activated documents are, the more the course deals with a large span of content.

Insert Tables 1 & 2 about here

7. **Applications of the Method**

The method presented above has two main purposes. Firstly, we plan to integrate this method into *Apex*, an Intelligent Tutoring System that automatically assesses course summaries (Dessus, Lemaire, & Vernier, 2000; Lemaire & Dessus, 2001). *Apex* takes as input a student essay and returns an assessment based on a semantic comparison with the relevant parts of the course, performed by LSA. For each relevant unit of the

course, the system informs the student about the way the content is covered. For instance, the system could say "*You have covered very well the section about X. You have covered badly the section about Y*". Then the student is asked to work on the essay once more. Assessments at the paragraph level were delivered as well as a general grade. Our method will be integrated into *Apex* as a way to link by hypertext each content unit to other related readings. At each moment students can perform three tasks: read the course; read the readings proposed by our method; write their unit course summary. Once integrated into *Apex*, this method could help teachers at a distance in two important tasks: a) help them to assess objectively a large amount of student essays; b) help them to deliver reading advices more connected to the comprehension and the interest of students.

Secondly, our method can also be used to verify that the course content adequately matches both teacher's intention (few topics covered, close enough to each other) and the students' prior knowledge. The method performs a semantic "scanning" of the course content and highlights the closer readings to the course. Then the teacher can assess whether the content of these readings is related to the course content and revise the course accordingly.

8. **Discussion**

We have shown that a solution to the problem "next text to be read" requires that the following steps be completed: 1°) simulate some aspects of discourse comprehension by an appropriate model (Kintsch, 1988); 2°) use a high-dimensional space to model the relations between units of knowledge delivered during the course; 3°) automatically create hypertext links between course content and readings related to the course. This method proposes readings related to a course without human pre-processing. However, the results presented above have to be confronted with an empirical evidence. We plan to perform a test in which students would be confronted to a course and the readings retrieved by the method. Then, they would be given several tests, whose results would be compared with data processed by LSA. Finally, the students would be asked to judge the adequacy of the documents retrieved by the method.

However, two shortcomings remain to be addressed. Firstly, the simulation of the comprehension process refers to a *generic* student model rather than a *specific* one. A new version of *Apex* (Dessus & Lemaire, 2002) that incorporates a student model, will address this problem. Secondly, the simulation refers to an integration of the course knowledge into the student long-term memory. Another experiment is planned in which a "window" slides as the course is delivered. This way, only the last *n* paragraphs of the course are subject to activation (i.e., are represented in the student' short-term memory). This could lead to a more reliable simulation of students' course understanding.

One of the assumptions of LSA's authors is that a consequent amount of knowledge (if not all) can be represented by a text input without important loss. This assumption may lead to a reductionist view of knowledge acquisition and representation, which is criticized by some researchers (e.g., Glenberg & Robertson, 2000;

Perfetti, 1998). Although such co-occurrence-based mechanisms (i.e., distributional approaches, see Redington & Chater, 1998) processed by LSA do not take into account all aspects of knowledge acquisition or representation, empirical research showed its closeness with human's one. We are aware of the idea that, if LSA can mimic some aspects of human comprehension, it is based upon theoretical assumptions that are disputed in the literature. For instance, Gernsbacher (1990) developed a concurrent model of language comprehension that is iterative but partially cumulative, depending on the way the content of a text is adequately understood in the reading of its initial sentences (called "foundation").

More precisely, the method presented above addresses some current problems about Intelligent Tutoring Systems. The type of interface appropriate to learning is often discussed by considering the following opposed arguments: —provide users with intelligent interfaces by allowing rich, although rather opaque actions (Gentner & Nielsen, 1996); —provide users with metaphor-based interfaces by allowing transparent, although rather simple actions (Shneiderman, 1992). We claim that our method allows both transparent manipulations through automatic hypertext linking and complex processing through the use of LSA.

Acknowledgements

We wish to thank Gerhard Dalenoort, Benoît Lemaire, Erica de Vries and two anonymous reviewers for thoughtful comments on a previous version of this paper; Pascal Bressoux for providing the course; and Jacky Beillerot for providing the reference textbase.

References

Blustein, J., & Webber, R. E. (1995), Using LSI to evaluate the quality of hypertext links*, Paper presented at the ACM SIGIR IR and Automatic Construction of Hypermedia: a research workshop*.

Brusilovsky, P. (1996), Methods and techniques of adaptive hypermedia, *User Modeling and User Adapted Interaction* ***6***(2-3), 87-129.

de Bra, P., Brusilovsky, P., & Houben, G.-J. (1999), Adaptive hypermedia: From systems to framework, *ACM Computing Surveys* ***31***(4).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990), Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science* ***41***(6), 391-407.

Dessus, P. (2000), Construction de connaissances par exposition à un cours avec LSA [Simulating knowledge building with LSA through lecture following], *In Cognito* ***18***, 27-34.

Dessus, P., & Lemaire, B. (2002), Using production to assess learning: an ILE that fosters Self-Regulated Learning, In S. A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Intelligent Tutoring Systems ITS-2002* (pp. 772-781). Berlin: Springer.

Dessus, P., Lemaire, B., & Vernier, A. (2000), Free-text assessment in a virtual campus, In K. Zreik (Ed.), *Proc. International Conference on Human System Learning (CAPS'3)* (pp. 61-76), Paris: Europia.

Gentner, D., & Nielsen, J. (1996), The anti-Mac interface, *Communications of the ACM* ***39***(8), 70-82.

Gernsbacher, M. A. (1990), *Language comprehension as structure building*, Hillsdale: Erlbaum.

Glenberg, A. M., & Robertson, D. A. (2000), Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning, *Journal of Memory and Language* ***43***(3), 379-401.

Goffinet, L., & Noirhomme-Fraiture, M. (1996), A*utomatic hypertext link generation based on similarity measures between documents*, Namur (Belgium): Computer Science Institute, Research Report FUNDP.

Kintsch, W, (1988). The role of knowledge in discourse comprehension: A Construction-Integration model, *Psychological Review* ***95*** (2), 163-182.

Kintsch, W. (2001), Predication, *Cognitive Science* ***25*** (4), 173-202.

Landauer, T. K., & 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. K., Egan, D., Remde, J., Lesk, M., Lochbaum, C., & Ketchum, D. (1993), Enhancing the usability of text through computer delivery and formative evaluation: the SuperBook project, In C. McKnight, A. Dillon, & J. Richardson (Eds.), *Hypertext, a psychological perspective* (pp. 71-136), Chichester: Ellis Horwood.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998), An introduction to Latent Semantic Analysis, *Discourse Processes* ***25***(2-3), 259-284.

Lemaire, B., & Dessus, P. (2001), A system to assess the semantic content of student essays, *Journal of Educational Computing Research* ***24***(3), 305-320.

Lemaire, B., & Dessus, P. (2003), Modèles cognitifs issus de l'Analyse de la sémantique latente [Cognitive modelling with LSA], *Cahiers Romans de Sciences Cognitives* ***1***(1), 55-74.

Payne, S. J. (1991), Interface problems and interface resources, In J. M. Carroll (Ed.), *Designing Interaction* (pp. 128-153), Cambridge: Cambridge University Press.

Perfetti, C. A. (1998), The limits of co-occurrence: Tools and theories in language research, *Discourse Processes* ***25***(2-3), 363-377.

Quint, V., & Vatton, I. (1992), *Hypertext aspects of the Grif structured editor: Design and applications,* Rocquencourt: INRIA Research Report #1734.

Redington, M., & Chater, N. (1998), Connectionist and statistical approaches to language acquisition: A distributional perspective, *Language and Cognitive Processes* ***13***(2/3), 29-91.

Rosé, C. P., Bhembe, D., Roque, A., Siler, S., Srivastava, R., & VanLehn, K. (2002), A hybrid language understanding approach for robust selection of tutoring goals, In S.

A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Intelligent Tutoring Systems ITS-2002* (pp. 552-561), Berlin: Springer.

Salton, G., & McGill, M. J. (1983), *Introduction to modern information retrieval*, New York: McGraw-Hill.

Shneiderman, B. (1992), *Designing the user interface*, Reading: Addison-Wesley.

Turney, P. D. (2001), Mining the web for synonyms: PMI-IR versus LSA on TOEFL*, 12th European Conference on Machine Learning (ECML-01),* Fribourg (Germany).

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P., Kintsch, W., & Landauer, T. K. (1998), Learning from text: Matching readers and texts by Latent Semantic Analysis, *Discourse Processes* ***25***(2-3), 309-336.

Philippe Dessus, Laboratoire des Sciences de l'éducation
1251, av. Centrale, BP 47, Université Pierre-Mendès-France, 38040 Grenoble CEDEX 9 France
E-mail: Philippe.Dessus@upmf-grenoble.fr

==========

Figure 1

Activated documents as a function of the amount of course paragraphs computed by LSA (see description of step 2 above)

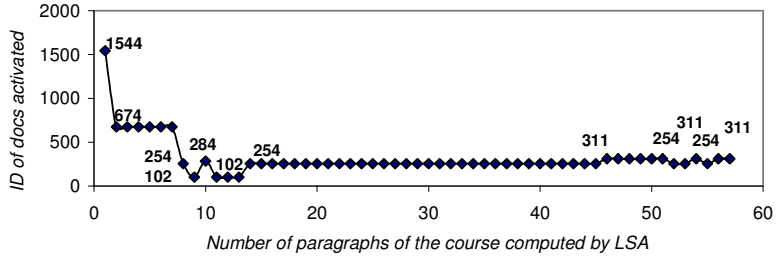

Table I
List of titles of the documents activated (short readings) and the amount of large readings retrieved (into brackets), once the course is delivered. In italics, query performed to retrieve documents presented in Table II below

ID of the last two activated documents (Short readings)	Results of the Internet Query (large readings), each word being added to the previous words to compose a new query (into brackets, number of documents retrieved)
Doc #311: *Research for Improving Quality of Instruction*	Doc #311 title words: research improving [1 121 040] quality [598 771] instruction [84 851]
Doc #254: *Learning Militancy in Youth Organizations*	Doc #254 title words: learning [65 248] militancy [156] youth [117] organizations [96]

Table II

List of the first ten large readings retrieved, once the entire course is delivered. Query performed the 20th january, 2002, using the Alltheweb search engine. In boldface, readings relevant to the course content

Large readings titles. In boldface, titles relevant to the course content
1. Research and collection of the Royal British Columbia Museum at Victoria, Canada
2. A large web page containing news about education (EducationNews.org site)
3. Vita of an Educational Researcher at University of North Carolina, Chapel Hill, USA
4. Text of a project about pedagogical militancy, Inter-American council for Integral development, USA
5. List of research projects, Educational dep/Political Sciences dep, Brown University, USA
6. Cold-War International History Project Bulletin
7. "School Characteristics and Educational Outcomes", paper published in the *Educational Administration Quarterly* review
8. Kellogg Foundation Publication about family and child caring, USA
9. "Redefining the politics over aboriginal language renewal, paper published in *The Canadian Journal of Native Studies*, Canada
10. Intro to Afro-American Studies, Murchison Community Center, Ohio, USA