RECOMMENDING AND SELECTING APPROPRIATE RESOURCES DURING ON-LINE PROBLEM SOLVING

Gregory A. Krudysz and James H. McClellan
Department of Electrical Engineering
Georgia Institute of Technology
krudysz@ece.gatech.edu jim.mcclellan@gatech.edu

Abstract

In this paper, we present a model of a personalized tutoring agent derived from a set of preliminary student data that was acquired using a web-based question and answering system. The system, called ITS, has been deployed in a second-year Signal Processing ECE course as an on-line supplement to the traditional homework. ITS allows students the opportunity to practice answering concept-centric questions and tests students' conceptual understanding through instructor assigned questions. The long term research goal is to develop an interactive learning environment that provides personalized tutoring via a set of questions and web-based content. Through a data-driven approach, we apply Hierarchical Bayesian models to develop a probabilistic conceptual framework for establishing and tracking the conceptual state and growth of students as they interact with questions and course related resources. We present preliminary results from our system, and discuss ITS in the context of extensions to account for conceptual correlations, a priori labeling, and temporal prediction.

Introduction

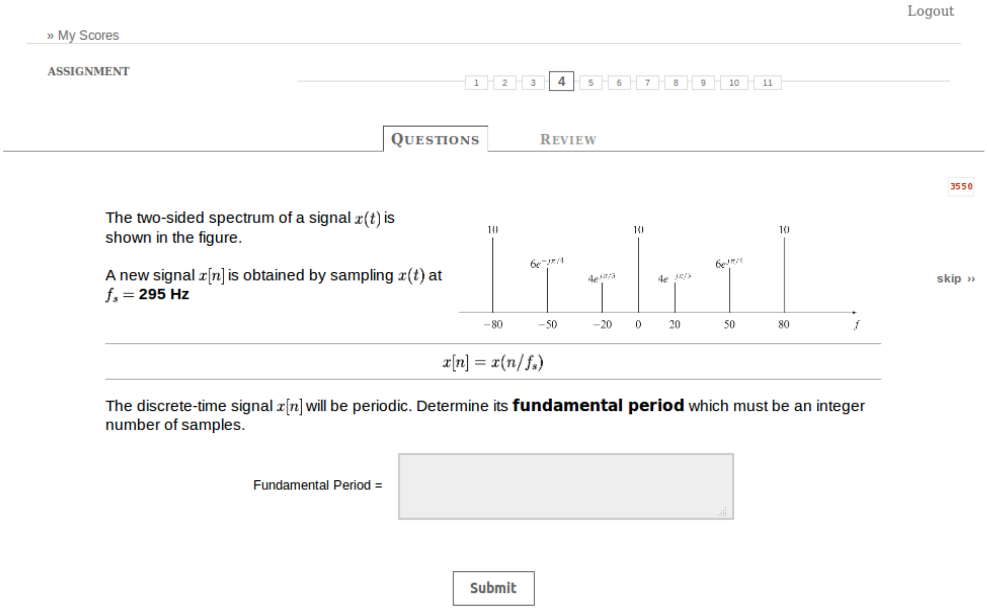
Tutoring is a form of assisted learning whose primary purpose is to enable conceptual growth. There are two main challenges facing developers of computer-based tutors. First, and perhaps the most equivocal task is in constructing a proper *mental model* of the user. The system simply may not have enough information to establish what the learner *knows* and more importantly what the learner *needs* to know and *when* a tutor's assistance is best served. In addition, each student is equipped with a varying background of experiences, possibly filled with various gaps in knowledge and misconceptions. The second challenge pertains to proper *knowledge representation* within the system. What knowledge constructs are required and to what degree they can help during tutee's impasse is highly contextual on the conceptual understanding of the student and the availability of a relevant resource within the system. There is a strong *inter*-relation between the *intent* of the user and the ability of the tutor to *infer* user's present state, and the *intra*-relation governing the tutee's and the tutor's capacity to advance to the goal state. This can be summarized by the reachability *criteria* which posits the likelihood of reaching from one state to another in the presence of enabling resources.

Copyright @ 2014, American Society for Engineering Education

In this paper, we develop a data-driven approach to education by describing the *mental model* as a structure consisting of student's *beliefs* and the process of learning as *belief growth* or *belief revision*. We seek to identifying belief structures which can represent how students reason in the context of external knowledge. “Students do not reason with all their beliefs but only with the subset which they are 'aware' is relevant.” [1] Thus, the central problem is to identify conceptual structures which could be relevant to a student in order to facilitate appropriate knowledge transfer during problem-solving activity. On the system side, accurate knowledge representation is vital for predicting what the student believes and what is necessary for the conceptual change to flourish.

ITS

Through our data-mining efforts, we develop a model of the tutoring agent by deriving a statistical relationship between *students*, *questions*, and system based *resources*. The aim of the agent is to assist users by recommending and selecting an appropriate resource during a problem-solving activity. From our database consisting of 400,000+ question records, and over 5,000 resource items, we model students according to their ability, questions according to difficulty, and resources according to their conceptual relevancy. We establish a set of relationships between students, questions, and resources by building statistical profiles based on their common underlying conceptual characteristics. These concept-based profiles will serve as models for tailoring user-directed feedback, such that student activity can be channeled toward refined learning strategies to enable improved problem solving concept comprehension through appropriate resource selection. Specifically, we focus on identifying user actions that would enable students to form useful associations of concepts with related resources in order to enhance their conceptual understanding and growth. A student view of the ITS system [2] is shown in Fig. 1.



The two-sided spectrum of a signal $x(t)$ is shown in the figure.

A new signal $x[n]$ is obtained by sampling $x(t)$ at $f_s =$ **295 Hz**

$$x[n] = x(n/f_s)$$

The discrete-time signal $x[n]$ will be periodic. Determine its **fundamental period** which must be an integer number of samples.

Fundamental Period =

Fig. 1: Student Interface of the ITS system

Bayesian Perspective

Over the years, a number of modalities have been proposed in an attempt to model a learner's problem solving activity. This task is formidable because problem solving is highly dependent on specific knowledge domain and its expertise requires the acquisition of content based skills. In a tutoring system, content that is meaningfully represented and highly structured can be very helpful in answering questions that users may encounter in the problem domain. There is a strong relationship between the roadblocks that learners encounter and the types of supporting resources that are available to enable them to overcome the obstacles [3], [4].

Prevailing web-based systems structure their content as lists or hierarchical maps and attempt to link these resources with a set of meta-data keywords, also known as tags. This process is often error prone and inefficient as it requires experts to individually label vast data repositories with tags while maintaining a proper ontological sequence and uniformity across topics of interest. It is often difficult to decide how many tags ought to describe a particular item of content and whether tags should be assigned appropriate weights to denote saliency. Tag maintenance is also burdensome, since the addition of new tags may lead to biased data labeling, while deletions or relabeling can lead to confusion and unlabeled content.

Early AI researches sought to model human behavior with the intent of simulating the learning process. In the 1960's, Herbert Simon, one of the AI pioneers, proposed that "each individual would be described by a different program, and those aspect of human problem-solving that are not idiosyncratic would emerge as the common structure and content of the programs of many individuals" [5]. Although early attempts at modeling individual users and their cognitive "structures" were based on elaborate symbolic logic and iterative procedures, recent research on human subjects [6],[7] have advocated probabilistic approaches which appear to conform to the way our minds construct knowledge through the act of observation. In this perspective, the human mind relies empirically on relevant cues and features within our environment in order to construct representational knowledge *structures* from the data of our everyday experience.

Human reasoning could be modeled by Bayesian framework, where probabilistic interpretations are based on prior experience and our understanding is updated with the available data. That is, we begin with a *hypothesis*, or a conceptual structure, that defines our initial understanding, but is subject to change with our ability to identify evidence which corresponds to the hypothesis. "Bayesian principles dictate how rational agents should update their beliefs in light of new data, based on a set of assumptions about the nature of the problem at hand and the prior knowledge possessed by the agents" [6]. Formally, the Bayes Theorem is expressed as:

$$P(\mathbf{h}|D) = P(D|\mathbf{h})\ P(\mathbf{h})$$

$$\text{posterior} = (\text{likelihood}) \times (\text{prior})$$

where **D** stands for observational data and **h** is the hypothesis or assumption. Thus, if the initial (prior) belief is weak, then with the addition of supporting evidence of high likelihood, the strength of the concluding belief increases.

In this paper, we are guided by the Bayesian probabilistic inference principles to design a tutoring system capable of learning from data representations of both the learner and his/her conceptual environment. We adapt Hierarchical Bayesian Models which infer a set of parametric structures governing how concepts *inter* – connect across a collection of resources. “Embedded in a hierarchical Bayesian framework, this approach can discover the correct forms of structure (the grammars) for many real-world domains, along with the best structure (the graph) of the appropriate form” [7]. In the context of our system, the structure can be best understood as a set of distributions across concepts. Hence, a learner is represented by a model of a mixture of underlying set of concepts. Likewise, each resource item in the repository is also modeled as a mixture of concepts, which are inferred from the content of that resource.

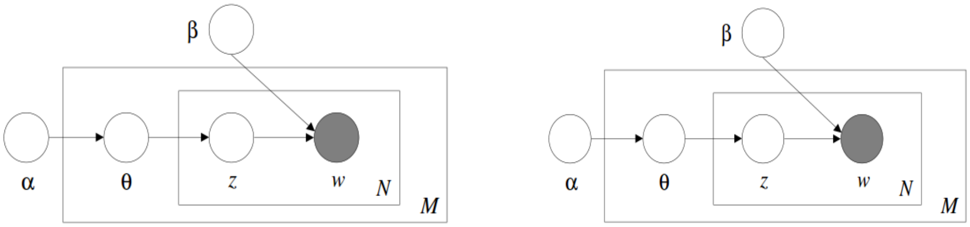
Specifically, we model *learners* by the resources that they observe or interact with; and we model the learning environment by the collection of resources that are available throughout the system. The *intra* – connection between the learner model and the available knowledge repository is expressed through unobserved or latent mixtures components, which for our purposes are representations of concepts forming a conceptual belief structure.

Model

The building block of the Bayesian probabilistic model is the *latent Dirichlet Allocation* (LDA) model [8], [9], which in the context of text processing attempts to infer a set of hidden (*latent*) topics from a collection of text documents and the words contained within them. LDA is a generative model, which posits that each *word* within a document is *generated* from a randomly chosen topic. This generative formulation presupposes a context on each of the words within a given text document. One of the strengths of generative probabilistic models is their ability to discriminate among contexts, thus allowing for these models to learn the meaning and usage of words. For instance, topics generated from these models are able to account for synonyms and polysems [10]. Synonyms are different words which are semantically related, for example: *lucky – fortunate*. Polysems are words that can have different meanings based on context, for example: *book*, as in a collection of pages, or “*book* a trip”. Context allows learners to make proper judgment as to what is relevant and thus enables for an appropriate selection of resources.

In a LDA model, topics are themselves generated from a mixture of topics which are represented by a *Dirichlet* distribution. These distributions are assumed to be specified by a parameter, which remains constant for each document. This hierarchical structure establishes a dependency relationship between the words in a document, latent topics, and documents. A graphical model for LDA is shown in Fig. 2 (a), where **w** represents the observed word (shaded), **z** is a topic, and **θ** is the topic mixture model describing topic proportions within each document. The model uses a plate notation to denote stacking of **M** documents, with a collection of **N** words within each document. The important part of this model is the fact that statistically the most plausible topic is selected to represent each word and that each topic is selected from the most plausible mixture of topics, based on the collection of documents. The

Bayesian model statistically combines words, topics, and documents with the inherent goal of discovering the most probable relationship among these objects. LDA infers the unknown topic structures by choosing the set of assumptions and beliefs expressed in the documents.



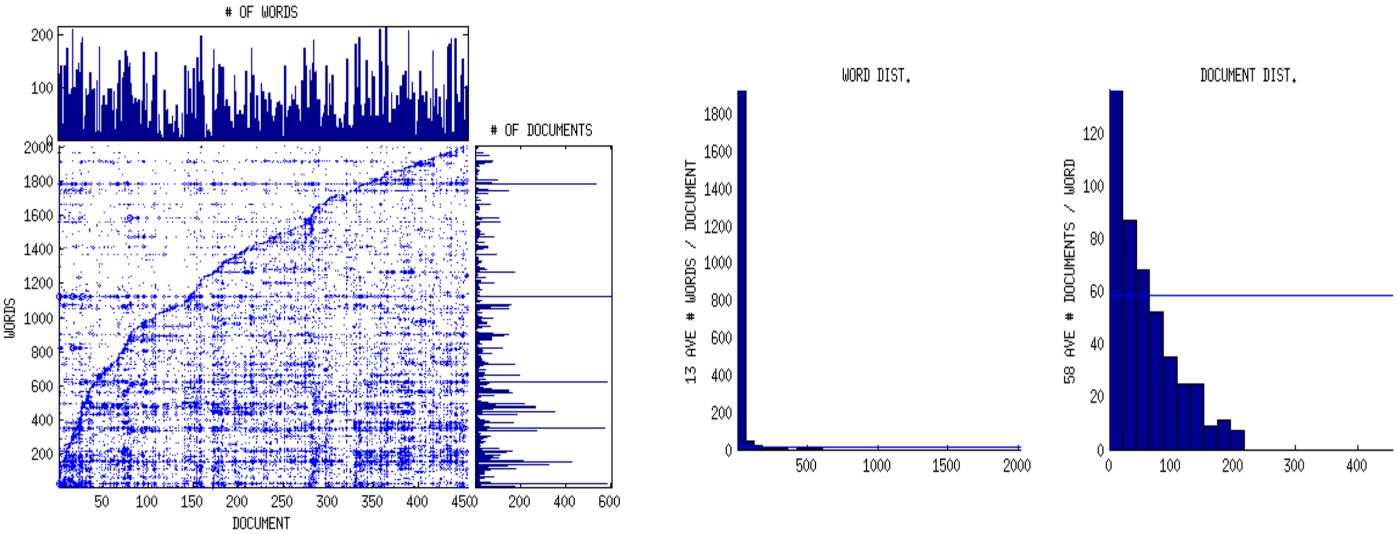
(a) (b)

Fig. 2: Plate notation for (a) LDA graphical model (b) ITS – LDA based model

Given its data-driven nature and sound statistical foundation of the hierarchical Bayesian models, we apply the LDA framework to establish specific knowledge domain relationships within the ITS. Our goal is to learn the underlying conceptual structure characterizing system resources which allows for the discovery of meaning and its conceptual applicability. We propose to divide the knowledge domain of the ITS into two repositories: one consisting of a bank of questions and another containing supporting knowledge domain resources. Each repository is described by a separate LDA model which captures the underlying conceptual structure. This model is shown in Fig. 2 (b), where $\mathbf{Q_z}$ and $\mathbf{R_z}$ represent concepts associated with the *Question* and *Resource* repository. In the *Question – LDA* model, there are **M** questions which are governed by a mixing set of concepts characterized by $\mathbf{Q_\theta}$. Each word **w** within the question text is randomly drawn from the mixing concept distribution. Similarly, the *Resource – LDA* model describes a collection of **M** text-based resources from which concepts $\mathbf{R_z}$ are derived.

Datasets

The *Question* database is a collection of 542 question documents, of which there are 337 multiple-choice, 78 matching, and 127 computed questions. After preprocessing and filtering out stop-words along with words which are less than three characters long, the size of the vocabulary is reduced to just 730 words. The *Resource* repository has been created by scanning the course textbook into a database. The book has been atomized, whereby each fragment of the book has been separately stored and labeled with either *chapter*, *section*, *paragraph*, *image*, *equation*, or *index-term* tag. In this study, we represent each document by a paragraph from the book, with the intent of recommending a pertinent paragraph as a hint or a study aide during problem-solving activity. We use the first four chapters of the book to obtain 455 paragraphs, and a vocabulary of 2015 words. A *co-occurrence* matrix for the *Resource* database is shown in Fig. 3, where in (a) the count of words and documents are shown. The word and document distributions are shown in part (b), where the horizontal line across the histogram denotes the average. Note that on average there are 13 words per paragraph, and on average there are 58 documents associated with each word.



(a) **(b)**

Fig. 3: *Resource* dataset (a) document-word *co-occurrence* matrix (b) word and document distributions.

Results

LDA models produce a set of topics that are represented by a mixture of observed words. In the ITS framework, each topic is a concept derived from the knowledge domain. Thus, a concept can describe a question or a supporting resource, and also it can serve as a context or a predictor for each word within the corpus. In topic models, as is often the case with clustering algorithms, the number of sought after topics is generally unknown, yet it must be specified. In this study, the number of concepts of interest is set to 50. Fig. 4 illustrates four of the most salient concepts from the *Question* database, where each concept is represented by a distribution across words. In each of the distributions, only the top ten most plausible words are shown, however, each concept is expressed by the proportional contribution of each of the words. Similar results are presented in Fig. 5 for the paragraph based documents. It is worth noting that the concepts derived from both of the datasets are different, as each collection was modeled separately by a distinct LDA model. Nevertheless, some of the words do overlap, and thus some probabilistic assumptions are conceptually stronger than others.

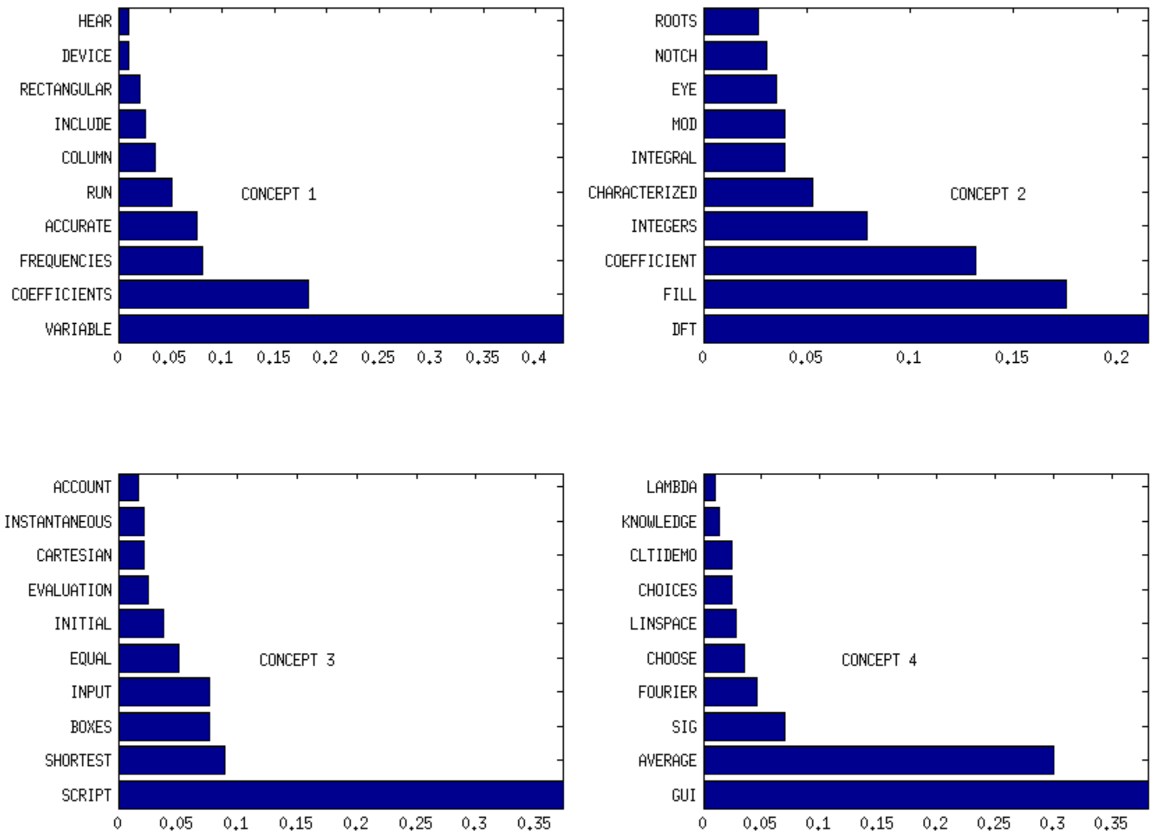


Fig. 4: LDA on the *Question* data set. Top four produced concepts with 10 most probable words.

A general framework for the ITS system can be further extended to establish a relationship between the *student*, a set of *questions* and a repository of *resources*. This scheme is illustrated in Fig. 6, where each component is characterized by the LDA model with its own corresponding conceptual foundation. The *student* model is generated from the observed content. In ITS, student activity is recorded, capturing the sequence of questions and resources that are accessed, along with vital statistics, such as scores, timing, and dwell duration. In addition, users are encouraged to rate the difficulty of questions and specify the academic significance of their problem-solving sessions. In the absence of prior data, concepts are uniformly represented with each mixing word assigned to a low probability. The *question* and *resource* structures are analogous to the *question*-*resource* LDA model describe in Fig. 2.

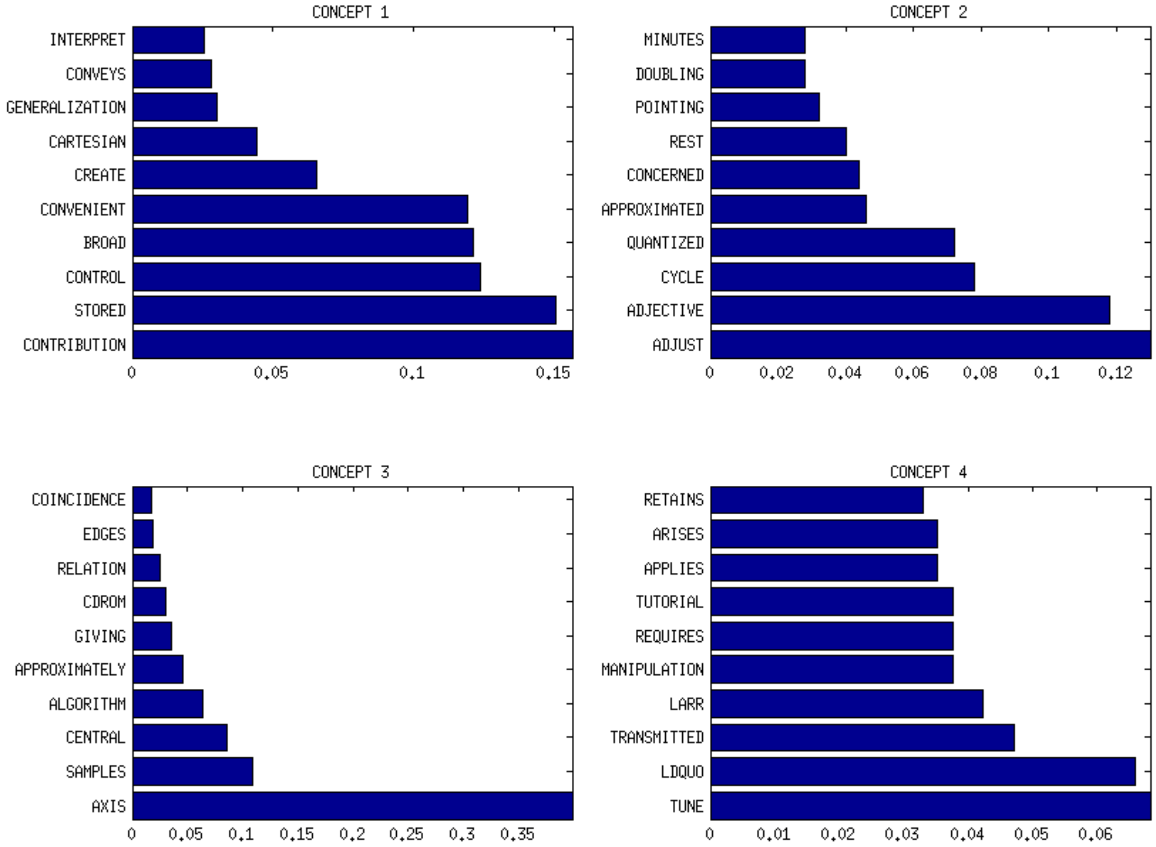


Fig. 5: LDA on the book *paragraphs*. Top four produced concepts with 10 most probable words.

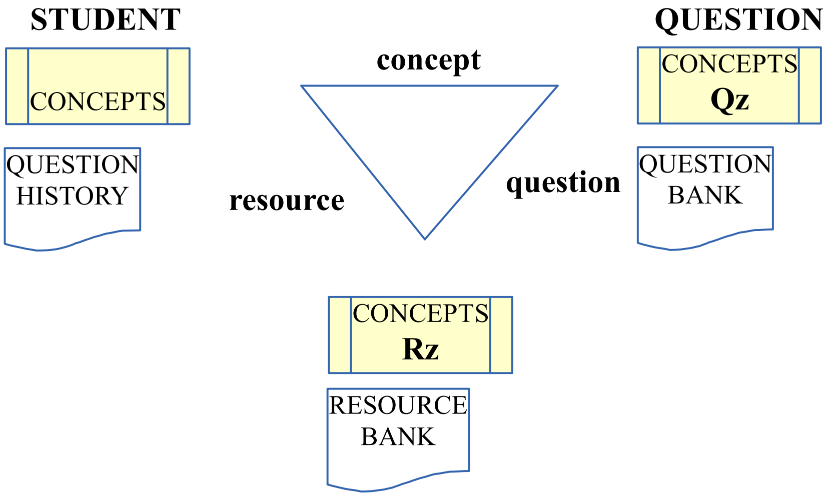


Fig. 6: ITS framework – three concept LDA models for *student*, *question*, and *resource* structures

Extensions

Hierarchical Bayesian models have been extended to account for some of the limitations associated with the LDA framework. We outline some of the proposed modifications to the LDA model and discuss their applicability to a tutoring environment.

At a fundamental level, the document-word relationship performs clustering to establish a similarity between documents and words. “Two words are similar to the extent that they appear in the same topics, and two documents are similar to the extent that the same topics appear in those documents” [11]. Although two conceptual structures can be compared to determine their similarity, LDA is unable to correlate between topics. This limitation stems from the fact that each topic mixture is generated independently from the other mixtures, as stipulated by the Dirichlet distribution. To account for the possibility that the presence of one topic within a document might be correlated with another topic, another hidden variable can be introduced to capture this relation. In Fig. 7, a graphical model of the Correlated Topic model (CTM) is shown, where the **β** parameter attempts to capture the correlation among all of the topic distributions. The primary advantage of the CTM is attributed to its predictive ability to associate one item with another based on the correlations of their *generating* topics. Thus, CTM offers an intuitive mechanism for recommending one resource based on an already observed one during a tutoring session.

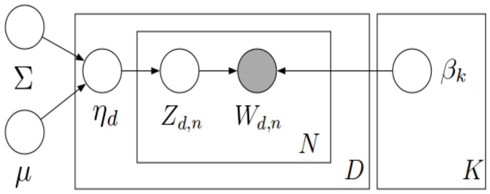


Fig. 7: Correlated Topic Model (CTM)

Any model which has only one source of information is limited in its scope. In Fig. 8, the Author Topic Model (ATM) [12] is shown, which builds upon the CTM by introducing additional observation within each document. The *author* label $\mathbf{a_d}$ is added and another distribution **θ** is modeled to represent the relationship between *authors* and topics. In fact, this label is a convenient description for any additional information that the system may have about each document, as the topic mixtures have a known generative source $\mathbf{a_d}$. In the tutoring context, this label is simply a tag, which has been associated with the user log, a question, or a supplemental resource. It can also embody a weight given in support of a particular topic distribution. Thus for the learner's *Question History*, it can denote user ability as derived from the question scores. User difficulty ratings can provide additional information on the significance and quality of the *Question Bank*. Prior knowledge on the ordering of content within the *Resource Bank*, such as section or chapter number add suitable chronology to the system. Again, the advantage of AT model is to bolster the meaning and usefulness of the conceptual structure embedded within the system.

Learning is a sequential process, whereby knowledge is encoded and refined through extensive practice. Comprehension is based on the number of opportunities available for rehearsal, the intervals between study sessions, frequency of practice, and the contextual associations formed between the new and the existing knowledge. “Speed and probability of accessing a memory is determined by its level of activation, which in turn is determined by how frequently and how recently we have used the memory” [13]. Dynamic topic models (DTM) [14], add the time component to the inference model by describing the evolution of topics over sequentially organized collection of documents. In Fig. 9, a graphical model is shown with each time slice represented by the LDA model, where α and β parameters capture the distributions of the dynamics of the underlying topics. This generative model subsumes that global topics evolve smoothly, yet are adept at capturing topic progressions between successive document observations. A tutoring system that tracks and fosters conceptual growth with a time-based model is better able to capture the evolving conceptual structure for each learner.

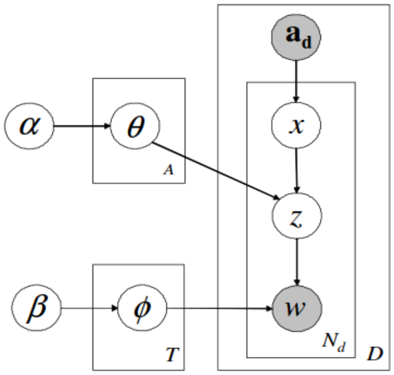


Fig. 8: Author Topic Model (ATM)

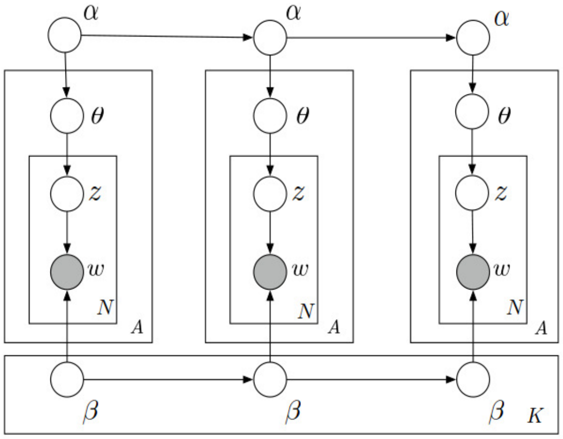


Fig. 9: Dynamic Topic Model (DTM)

Conclusion

In this paper, we have proposed a model for an interactive tutor grounded in data-driven probabilistic framework which conforms to the way humans reason in the uncertain world with limited observations. We have proposed a model describing the *intra*-relationships between a learner's conceptual state, a set of questions challenging student's implicit beliefs, and the conceptual knowledge representation afforded by the system. Our future work will extend this preliminary modeling to incorporate the CTM, ATM, and DTM models into our system design.

Acknowledgement

ITS development has been supported in part by the National Science Foundation Award No. 1041343 “Collaborative Research: CI-Team Implementation Project: The Signal Processing Education Network”.

References

[1] J. Self, “Theoretical Foundations for Intelligent Tutoring Systems”, AAI/AI-ED Technical Report , Journal of Artificial Intelligence in Education, 1(4), 3-14, 1990

[2] G. A. Krudysz and J. H. McClellan, "Signal processing education through concept discovery and resource selection practice”, ICASSP 2013.

[3] K. VanLehn, S. Siler, C. Murray, T. Yamauchi, W. B. Baggett, “Why Do Only Some Events Cause Learning During Human Tutoring?”, Cognition and Instruction, Vol. 21, Iss. 3, 2003

[4] William Bechtel, George Graham, “A Companion to Cognitive Science”, Vol 63 of Blackwell Companions to Philosophy, Wiley, 1999

[5] A. Newell and H.A. Simon, “GPS: A Program that Simulates Human Thought”, In Feigenbaum, E.A. & Feldman, J. (eds.), Computers and Thought, 279-293, 1963.

[6]. T. L. Griffiths, C. Kemp and J. B. Tenenbaum, “Bayesian models of cognition”. In R. Sun (Ed.), Cambridge handbook of computational cognitive modeling, Cambridge: Cambridge University.

[7] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, “How to Grow a Mind: Statistics, Structure, and Abstraction”, Science 11 March 2011: 331 (6022), 1279-1285

[8] Thomas Hofmann, "Probabilistic Latent Semantic Analysis", *Uncertainty in Artificial Intelligence, 1999.*

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, “Latent dirichlet allocation”, *J. Mach. Learn. Res.* 3 (March 2003), 993-1022.

[10] Steyvers, M., & Griffiths, T. “Probabilistic topic models”, In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis.* Hillsdale, NJ: Erlbaum.

[11] D. Blei and J. Lafferty, “Correlated Topic Models”, Advances In Neural Information Processing Systems, 18:147, 2006.

[12] Rosen-Zvi, M., Chemudugunta, C., Griffiths, T. L., Smyth, P., and Steyvers, M. “Learning author-topic models from text corpora”, *ACM Transactions on Information Systems, 28(1),* Article 4

[13] J. R. Anderson, “*Cognitive Psychology and Its Implications: Seventh Edition*”, New York: Worth Publishing, 2010

[14] D. M. Blei and J. D. Lafferty, “Dynamic topic models”, *Proceedings of the 23rd international conference on Machine learning* (ICML '06). ACM, New York, NY, USA, 113-120