A knowledge-tracing model of learning from a social tagging system

Peter Pirolli & Sanjay Kairam

User Modeling and User-Adapted Interaction
The Journal of Personalization Research

ISSN 0924-1868

User Model User-Adap Inter
DOI 10.1007/s11257-012-9132-1

Springer

User Model User-Adap Inter
DOI 10.1007/s11257-012-9132-1

ORIGINAL PAPER

A knowledge-tracing model of learning from a social tagging system

Peter Pirolli · Sanjay Kairam

Received: 1 November 2010 / Accepted in revised form: 17 October 2011

Abstract We propose a user model to support personalized learning paths through online material. Our approach is a variant of *student modeling* using the computer tutoring concept of *knowledge tracing*. Knowledge tracing involves representing the knowledge required to master a domain, and, from traces of online user behavior, diagnosing user knowledge states as a profile over those elements. The user model is induced from documents tagged by an expert in a social tagging system. Tags identified with "expertise" in a domain can be used to identify a corpus of domain documents. That corpus can be fed to an automated process that distills a topic model representation characteristic of the domain. As a learner navigates and reads online material, inferences can be made about the degree to which topics in the target domain have been learned. We validate this knowledge tracing approach against data from a social tagging study. As part of this evaluation, we match the predictions of the knowledge-tracing model to individual participant responses made to individual question items used to test domain knowledge.

Keywords Cognitive models · User models · Latent Dirichlet allocation · LDA · Topic models · SparTag.us · Social tagging · Social web

P. Pirolli (✉) · S. Kairam
Palo Alto Research Center, 3333 Coyote Hill Rd, Palo Alto, CA 94304, USA
e-mail: Pirolli@parc.com

S. Kairam
Department of Computer Science, Stanford University, 353 Serra Mall, Office 376, Stanford, CA 94305, USA
e-mail: skairam@cs.stanford.edu

Published online: 03 November 2012

1 Introduction

The Web has become a primary resource for individuals to learn, in a self-directed manner, about science, technology, and medicine, often with the aim of eventually solving significant problems in areas such as personal health or gaining competence in emerging technical fields (Fox and Fallows 2003; Fox and Jones 2009; Horrigan 2006; Lenhart 2009). Social web systems, in particular, from tagging systems to social networks have additionally afforded the opportunity for large groups of people to curate and collaborate around content. For the self-directed learner who seeks to master available expert knowledge, one problem would be finding such pools of online expertise within such systems, and a second would be the problem of focusing attention on things that need to be learned rather than things that the user already knows. Recent work (e.g., Noll et al. 2009) has started to focus on the problem of identifying "expert" users and "expert" sets of knowledge from the social Web. Here, we focus on developing an approach that we argue is on the path to solving the second problem of personalized learning paths through online material: We propose a variant of *student modeling* using the computer tutoring concept of *knowledge tracing* (Anderson et al. 1990; Corbett and Anderson 1995). Knowledge tracing involves representing the knowledge required to master a domain, and, from traces of online user behavior, diagnosing user knowledge states as a profile over those elements.

Personalization in the domain of education is epitomized by human one-on-one tutoring, in which a student's attention is guided through a course of content and practice based on a tutor's expert understanding of the subject matter and the tutor's understanding of the student's current state of learning. This extreme form of personalized learning can yield up to two standard deviations of improvement relative to less-personalized classroom instruction (Bloom 1984). Sophisticated computer-based tutors have been able to achieve levels of personalization and learning outcomes comparable to human tutors in some domains (Corbett 2001), but these require considerable knowledge engineering efforts (Aleven et al. 2009). In contrast, the Web can provide cheap access to knowledge, but offers very little in the way of guiding a learner's attention to material that best suits their current state of learning while simultaneously maximizing their progress to achieving expert command of the subject matter.

As noted above, a core element of what makes an Intelligent Tutoring System (ITS) "intelligent" is its ability to accurately diagnose students' knowledge and adapt instruction accordingly. Most of these student models have been hand-crafted by considerable person-hours of knowledge engineering effort, and the focus is often on the modeling of procedural cognitive skills (Corbett 2001). There are, however, approaches (e.g., Foltz et al. 1999; Kakkonen et al. 2005) that have automatically induced a semantic representation of a domain (e.g., introductory psychology) from a given corpus of documents, used that to diagnose student task performance (e.g., writing an essay), and provided feedback and recommendations that ultimately improved learning.

In this paper, we aim to develop a diagnostic student model that is induced from documents tagged by an expert in a social tagging system. Expert tags can be used to identify a corpus of domain material. That domain content can be fed to an automated process that distills a representation of the topics characteristic of the domain (Rosen-Zvi et al. 2004; Steyvers et al. 2006). As a learner navigates and reads online material,

inferences can be made about the degree to which topics in the target domain have been learned. We test this knowledge tracing approach against data from a social tagging study summarized in Nelson et al. (2009). As part of this evaluation, we match the predictions of the knowledge-tracing model to individual participant responses made to individual question items used to test domain knowledge in the Nelson et al. study. To do this, we couple the domain topic model with a variant of psychometric measurement techniques (Pirolli and Wilson 1998). More specifically, our approach to building a framework for modeling user learning within a novel subject domain includes the following tasks:

- *Building a Topic Model to Represent User Knowledge States*: We employ a topic modeling approach based on Latent Dirichlet Allocation (LDA) (Blei et al. 2003) in order to induce the latent topics inherent in the subject domain. We utilize a corpus consisting of documents (Web pages) drawn from a social tagging system as well as documents browsed by users. Knowledge of the domain can be represented as the possession of different degrees of latent ability with respect to each of these underlying latent topics.
- *Developing the Measurement Framework*: Given topical information about the domain and observed data from users, we demonstrate how we can construct a measurement model capable of inferring the users' knowledge profiles across topics. Specifically, we utilize data from Web browsing traces and from patterns of observed responses to test tasks as inputs into this model. We present variations on this model that incorporate differing assumptions about the factors underlying individual learning differences.
- *Testing the Framework*: Finally, we test the predictive power of the different model variations against the observed pre- and post-test data from a study on e-learning (Nelson et al. 2009) using the social annotation system SparTag.us (Hong et al. 2008). In this model-based analysis, we explore the relative power of these models in predicting individual learning gains and explaining the increased group performance of users of the SparTag.us system with access to the annotations of an expert 'friend'. By comparing variations of the model that do or do not include topical information, we demonstrate how the inclusion of this information increases the overall predictive power of the measurement framework.

The development of student models in ITSs has led to a deeper understanding of learning in a variety of domains (e.g., Anderson 1984, 1993). An ancillary purpose for the development of our knowledge tracing methodology has been to gain a deeper understanding of learning from social tagging systems. In such systems, tags produced by experts presumably provide more efficient navigation cues for learners to follow. One of our aims is to provide evidence to support such an hypothesis through analyses derived from application of the knowledge-tracing model.

2 Background

2.1 Social tagging

The rise and proliferation of social web systems has led to an explosion of content ripe for consumption. Photo-sharing sites such as Flickr offer instant access to liter-

ally billions of images (Champ 2009), and the video-sharing site YouTube currently claims to serve 3 billion videos each day (YouTube 2012). Social bookmarking sites, such as delicious, Digg, and StumbleUpon have become conduits for the sharing of particularly rich content by allowing users to annotate content across the Web for themselves and for others. The success and power of these sites stems, in part, from the fact that while the individual act of bookmarking a website or other resource is fairly simple, the aggregate value created by the community as a whole can be enormous.

One effect of the cooperative sharing of annotations, such as tags, is the emergence of *folksonomies* (Golder and Huberman 2006). Folksonomies are vocabularies of tag labels which evolve through consensus in social systems to be similar to the engineered keyword vocabularies used to categorize library books (Robu et al. 2009). In both emergent folksonomies and constructed library keyword vocabularies, there are implicit expectations that the organizational structures and labeling can enhance user navigation. Browsing tools such as Mr. Taggy (Kammerer et al. 2009) have demonstrated that folksonomies can be leveraged to help users navigate social web systems by providing feedback and context for user queries. This contextual information is especially helpful for users who are searching for information in a novel domain and unaware of which keywords to utilize in search, a problem commonly referred to as the "vocabulary problem" (Furnas et al. 1987).

While these annotations are useful on their own, another emerging byproduct of this collective activity is the network of interaction which is formed by users annotating common resources, using common annotations, and forming connections with one another. Unsurprisingly, recent approaches at navigating these systems have followed many of the same information retrieval and ranking paradigms which have seen success in web search. FolkRank (Hotho et al. 2006), for instance, extends the popular PageRank algorithm for ranking web pages based on link structure (Brin and Page 1998); it treats delicious as a triadic graph connecting users, resources, and tags and uses the structure of this graph to rank any of these elements against a user query. Other prior work in this area has built on the HITS algorithm (Kleinberg 1999), an approach developed around the same time as PageRank which identifies two types of important entities, *hubs* and *authorities*. Noll et al. (2009), for instance, ranks users according to expertise in a particular domain utilizing the SPEAR (SPamming-resistant Expertise Analysis and Ranking) algorithm which builds on HITS. Abel et al. (2009), use an variant called SocialHITS which extends the notion of hubs and authorities to users, tags, and resources.

The effectiveness of these graph-based approaches in identifying expert users or relevant content has been explored across a diverse set of social systems such as email networks (Campbell et al. 2003; Zhang and Ackerman 2005), Q&A Forums (Zhang et al. 2007), e-Commerce sites (Yin et al. 2009), and Social Networks (Gayo-Avello and Brenes 2010; Weng et al. 2010; Loizou and Dimitrova 2012). In conjunction with appropriate user models, such approaches can power what Gena et al. call *social adaptive* systems (Gena et al. 2012), or systems which combine user and social data to improve personalization. Together, the previous research in the area of mining these networks has demonstrated that there is great potential for harnessing a vast amount of content from such systems to support individual users on the web.

2.2 SparTag.us

The SparTag.us social annotation system is a tool that allows users to annotate and collect paragraphs of interest from pages on the Web. Annotations are made either through highlighting of content or through the Click2Tag interface which allows users to tag pages and paragraphs easily by clicking directly on the words being read, as illustrated in Fig. 1. Once pages have been annotated, SparTag.us also copies the annotated paragraphs into a system-created notebook, along with the URL of the page.

What makes the system social is the ability not only to view one's own notebook of annotations, but also to subscribe to the notebook of another user by designating that user as a 'friend'. Figure 2 shows a portion of what a user might see when looking at the notebook of a friend. In this figure, we see the annotated paragraphs displayed in the main window, with tags displayed underneath each paragraph to which they are

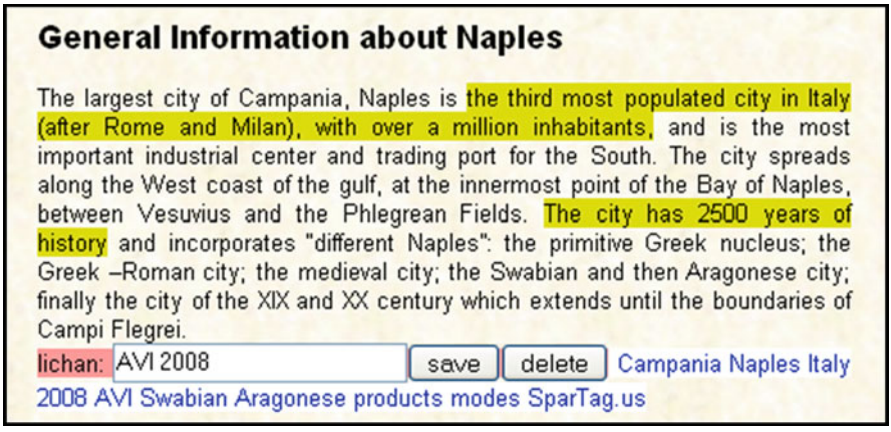

Fig. 1 SparTag.us allows for in situ highlighting and tagging of content, thus lowering the interaction costs for the production of annotations

Fig. 2 The SparTag.us notebook allows users to store annotated paragraphs as well as a tag cloud created from tags generated by the user

applied. In addition, on the right we see a tag cloud which displays the most common tags, with font size corresponding to the relative frequency of a given tag.

While SparTag.us was originally introduced as a low-effort system for tagging web content (Hong et al. 2008), these social features provided additional benefits by allowing for the sharing of expertise through annotations stored in notebooks. In prior study (Nelson et al. 2009), it was demonstrated that the presence of these social annotations when learning in a new domain had a significant effect on learning gains. In the following sub-sections, we will review the SparTag.us social reading experiment, describe how we applied a topic model to the content foraged and produced during the experiment, demonstrate how data from the experiment can be fitted to the variations of the proposed model, and evaluate the predictive power of these models. We will see that predictions generated by the measurement framework closely match observed data from the study and that insights generated from model predictions can help to explain differences in learning gains made by study participants.

2.3 The SparTag.us social reading study

Nelson et al. (2009) performed a study of SparTag.us aimed at testing the efficacy of the system. The study also tested the ancillary hypothesis that the availability of social tags produced by an expert would enhance self-directed learning. In and of itself, SparTag.us was not found to produce improved learning relative to commercial off-the-shelf tools (a standard browser and office suite configuration). However, learning was improved when a set of expert tags were made available through the SparTag.us interface.

2.3.1 Experimental conditions: contrasts of interfaces and the availability of expert annotations

The Nelson et al. (2009) study involved 18 participants solicited from various sources, including company interns and a local university. All participants had no previous experience with the SparTag.us tool, and six of the participants reported having some education in computing. The participants used Web resources to explore the complex domain of "Enterprise 2.0 Mashups" (roughly the intersection of "Enterprise 2.0" and "Web 2.0 Mashups". This choice of domain required participants to find and synthesize information from many web pages in order to learn enough to answer questions on the topic, as there was no single, centralized source of information on the domain area at the time of the study.

Participants were randomly assigned to three conditions:

- "*Without SparTag.us*" (**WS**), where participants had access to the Web and traditional note-taking tools such as a word processor or a pen and paper.
- "*SparTag.us Only*" (**SO**), where participants had access to the Web and the *SparTag.us* Web annotation system.
- "*SparTag.us with Friend*" (**SF**), where participants had access to the Web and to *SparTag.us*, as well as the annotations of an 'expert' friend named 'mjones', whose

notebook had been algorithmically constructed in order to provide pointers to key content.

These groups were intended to provide some simple contrasts (rather than a more complex factorial design experiment). The contrast of learning effects in SO v. WS groups would provide a test of improvements effected by the SparTag.us interface itself over the benchmark commercial off-the-shelf tools in the WS condition. The contrast of learning effects in the SF v. SO conditions would provide an indication of improvements effected by the availability of expert annotations in the SF condition over the control SO condition.

2.3.2 Learning tasks, procedures, and tests

Participants in all three groups were asked to find and read material in order to write reports on Enterprise 2.0 Mashups. The core study procedure involved: (1) a knowledge pretest, (2) learning in the domain area, (3) a knowledge posttest, and (4) essay writing. Participants were given a brief written statement of learning objectives, instructing them to read from any sources and take notes as they felt appropriate regarding the definitions, standards, benefits, issues, and examples relating to the topic area. Participants had one hour of unsupervised learning, with a break for lunch, and then 50 more minutes of learning. Sessions were logged, including URLs visited, content scrolling, and words written. The questions for the writing tasks and tests were solicited from experts, and additional detail about these and other aspects of the experimental setup and procedure is given in Nelson et al. (2009).

The pretest and posttest were designed to assess domain-specific knowledge about "enterprise 2.0 mashups." Two lists of 20 true-false questions were created, and each list was used as a pretest for half the participants in each group and as a posttest for the other half. The true-false questions were elicited from experts. Each list of 20 questions was designed to have an even distribution of easy and hard questions about enterprise mashups, as rated by 100 raters of varying expertise recruited via Amazon Mechanical Turk. Both tests were taken without access to tools or resources.

2.3.3 Summary of learning gains across conditions

Nelson et al. (2009) focused on a comparison of learning gains across the three groups (WS, SO, SF). For each subject, the learning gain was calculated according to the following metric:

$$Gain = \frac{\text{Posttest score} - \text{Pretest score}}{\text{Max score} - \text{Pretest score}}$$

Table 1 shows the main result of Nelson et al. (2009). An analysis of covariance was used to tease out the effects of non-experimental background variables correlated with learning. This analysis of covariance showed that the SF group had significantly greater gains than the SO group and the WS group. The difference in learning gains between the WS and SO groups was not found to be statistically significant. The details

Table 1 Mean learning gain scores in the Nelson et al. (2009) study of SparTag.us

	Learning gains	
	Mean	SD
SF	0.46	0.22
SO	0.13	0.32
WS	0.27	0.23

of the statistical contrasts can be found in the original Nelson et al. (2009) report. This analysis suggests that the SparTag.us interface itself provided no detectable impact on learning relative to the commercial-off-the-shelf tools, but there was a learning improvement effected by the availability of the expert annotations.

It should be noted that Nelson et al. (2009) reports groupwise statistical comparisons, and no detailed analysis of individual differences in background knowledge or learning were carried out. One aim of the current paper is to use a knowledge tracing model to drill down on the source of the learning gain effects observed in the groupwise comparisons.

2.4 Intelligent tutoring systems

As of the fall of 2008, over 4.6 million students were taking at least one online course, which represented a 17 % increase over the same period in the previous year (Allen and Seaman 2010), and this type of structured coursework represents only a small portion of the web-based learning that is occurring around the world. This growth speaks to the many benefits of web-based learning technologies over traditional pedagogical methods, some of which are detailed in Brusilovsky and Peylo (2003), such as the ability to facilitate learning for individuals without access to a classroom and to be accessed regardless of a user's computing platform. ITSs, in classroom situations, have enjoyed a great amount of success in improving learning within specific domains. By 2003, 1400 schools across the country were already using the Cognitive Tutor system developed by Carnegie Learning, and the company has documented many accounts of significant improvements in algebra education as a result (PACT 2005).

In order to successfully deliver and adapt content for a learner, an ITS typically uses three primary components: knowledge of the domain (*expert model*), knowledge of the learner (*student model*) and knowledge of teaching strategies (*tutor model*) (Hartley and Sleeman 1973). A wide range of knowledge representation techniques have been used measure knowledge and learning in expert and student models, including semantic networks (e.g., Carbonell 1970), case-based reasoning (e.g., Han et al. 2005), Bayesian networks (e.g., Conati et al. 2002), and production systems (e.g., Anderson et al. 1990). The ability to build accurate expert and student models and to adapt teaching strategies accordingly is the heart of what makes these systems "intelligent" (Shute and Psotka 1996). However, the creation of these models requires considerable person-hours of knowledge-engineering effort. It has been estimated that every hour of ITS instruction requires effort of about 200-300 hours of development time

(Murray 1999, 2003). More recently, authoring tools have been developed that have successfully reduced this ratio to 50-100 hours of development per hour of instruction (Aleven et al. 2009). For crucial, widely deployed courses (e.g., algebra) such level of effort to produce such high learning gains is quite cost-effective.

Of course, it would be prohibitive to develop an ITS for every nontrivial learning task facing users in their everyday lives, at home or at work. If we can obviate the need for this type of explicit knowledge-engineering effort and make it possible to generate self-tutoring systems for a wide variety of domains, there is great potential for these types of systems in helping people to educate themselves using the Web. Assuming that we have some sort of annotated corpus of documents containing knowledge of a particular domain, we need methods of automatically inducing student and expert models. Some approaches (e.g., Foltz et al. 1999; Kakkonen et al. 2005), to e-Learning applications have used Latent Semantic Analysis (LSA, Landauer and Dumais 1997) or Probabilistic Latent Semantic Analysis (PLSA, Hofman 1999) to induce the latent topics represented in a domain. LSA- and PLSA-based approaches are capable of automatically inducing a semantic representation of a domain from a curated corpus of documents and have been used in applications such as essay grading or summarization. We employ a similar approach based on *topic models* induced by LDA from social web resources.

2.5 LDA-based topic models

Our approach employs LDA (Blei et al. 2003) to induce an expert model of the topics in a domain. Each user's state of learning can be represented as a set of *topic ability parameters*. Each of these topic ability parameters is associated with a topic in the expert model and represents the degree of learning of that topic. Individual differences in efficiency of learning can be represented by *learning rate parameters*. These ability and learning rate parameters are induced from observed Web browsing behavior and, when available, responses to test questions.

The LDA topic model is a type of generative probability model that assumes that documents are produced as a probabilistic mixture of topics, where these topics are composed of probabilistic mixtures of words. Specifically, an LDA topic model is a type of three-level hierarchical Bayesian model (Blei et al. 2003). The LDA-based topic model assumes the existence of a latent structure, L, which represents the gist of a set of words (e.g. a document), g, as a probability distribution over some T topics. Each topic is, itself, a probability distribution over words, where words can be associated with one or more topics. The probability model for the ith word in a document conditional on the gist of the document can be specified as:

$$P\left(w_i|g\right) = \sum_{i=1}^{T} P\left(w_i|z_i\right) P\left(z_i|g\right) \tag{1}$$

where w_i is the probability that word w will occur at position i in a document corpus, and z_i is the topic of the ith word. The stochastic process for generating the ith word-token in a particular document involves selecting a topic z_i based on the

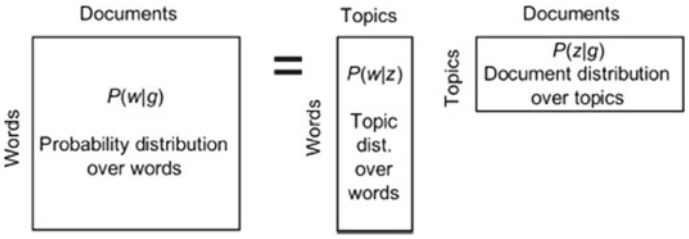

Fig. 3 A matrix representation of the Topic Model

conditional distribution of topics given the gist of the document, and then word w_i is selected based on the conditional distribution of words given topic z_i. Thus, in essence, $P(z_i|g)$ reflects the prevalence of the topic z_i within a document, and the conditional probability $P(w_i|z_i)$ is the prevalence of a word w_i within that topic.

Using the bag-of-words assumptions, the text can be re-arranged as a Word-Document co-occurrence matrix containing probabilities as in Fig. 3. This representation assumes a matrix with rows corresponding to a vocabulary of W (distinct) words in the collection, columns corresponding to documents, and individual cells containing the probability of a word occurring in a document. This resulting Word-Document matrix is assumed to be the product of a matrix of the probability of words within topics and a matrix of the probability of topics within documents. An observed Word-Document matrix is viewed as the result of stochastic draws using the probabilities depicted in Fig. 3.

3 Knowledge tracing based on topic models

3.1 Latent topics and user knowledge states

Probabilistic approaches to semantic representations, particularly those based on Bayesian approaches, have arisen as the result of rational analyses of cognition (Anderson 1990; Steyvers et al. 2006). Rational analysis is a framework in which it is heuristically assumed that human cognition adapts to the problems posed by the environment. Topic category judgments are viewed as prediction problems, and these can be guided by statistical inferences made from the structure of the environment, especially the linguistic environment. The structure of the topic model is motivated in part by consideration of the role of associative semantic memory in the formation of documents (Griffiths et al. 2007). Topic models have been used to predict many memory and linguistic phenomena in cognitive psychology. The statistical properties of the topic model representation have been shown not only to match the natural statistics of human language, but also to account for issues such as polysemy and asymmetry in ways that are hard to account for in spatial representations. In fact, the LDA-based topic model has been shown to out-perform LSA in predicting word association and a variety of other linguistic processing and memory tasks (Griffiths et al. 2007).

In our model, we utilize the topic model as a means of characterizing the underlying semantic spaces prevalent in a domain. Figure 4 presents a schematic representation of

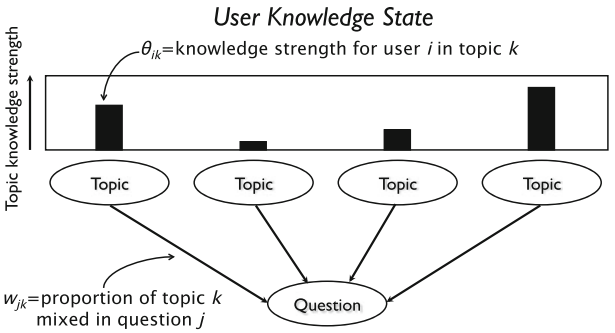

Fig. 4 A user knowledge state is represented as a profile over topics. A question tests a mix of topic knowledge

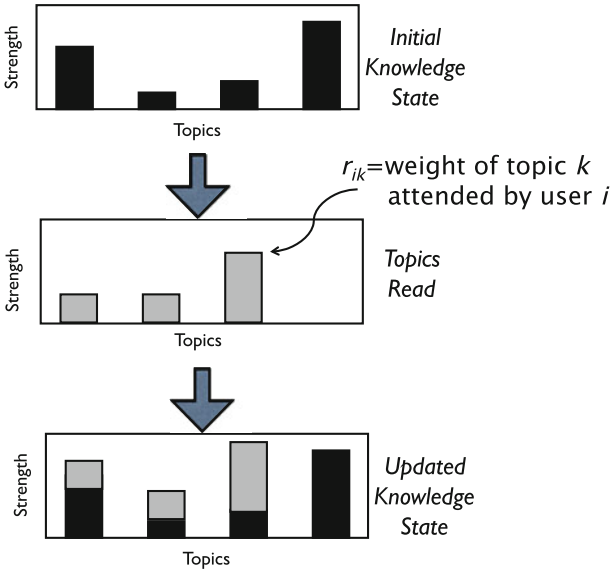

Fig. 5 Notional model of knowledge tracing based on Topic Modeling. A user's knowledge profile is modified by the mix of topics that they read about while browsing. This paper focuses on establishing the predictive power of a Topic Model fit to pretest measures of initial knowledge states, estimates of topic knowledge based upon online reading history, and posttest measures of updated knowledge states

the relationship between the knowledge state of a user and performance on a test item. It is assumed that each item requires that knowledge characterized by some mixture of topics be retrieved from the user's memory. This is represented in Fig. 4 by the links from the topic nodes to the question node. The knowledge state of a user, in turn, can be represented at any given point as different degrees of latent ability or strength with respect to the various topics, represented by the size of the bars in Fig. 4. The first task for the model is to induce an initial strength-of-knowledge profile from a user's observed responses to pre-test questions about the subject domain.

Figure 5 presents a schematic representation of assumptions about learning which are built into the model. It is assumed that each user starts the learning process with an initial knowledge profile, as described above, which represents his or her knowledge across topics. Reading about each of these topics while browsing the Web updates this knowledge profile. The task for the model is to induce the topics read by the user from the online traces of his or her individual browsing behavior and to relate this to other measurements of the changes in the user's knowledge state.

3.2 Item response models

Below, we report on an evaluation of a knowledge tracing model based on topic models induced by LDA. The evaluation was performed using a measurement model based on Item Response Theory (IRT). Item Response Theory provides a method for inferring a user's ability with respect to some area of knowledge from the user's pattern of responses to a set of test items (questions). The fundamental assumption in IRT is that given some latent trait θ being tested by a test item, the probability of a correct response to that item is a function of an individual's θ score (Kline 2005). While traditional IRT models assume that tests evaluate a single θ characteristic, it is often more realistic to interpret these test items as evaluating a mix of topics. The extension of these models to multiple dimensions has been demonstrated to be effective in accounting for simultaneous testing of multiple latent traits.

These multi-dimensional models can be split roughly into two types: *non-compensatory* models, where sufficient knowledge of all latent traits characterized by a test item is required to achieve a correct response, and compensatory models, where it is assumed that a high degree of ability with respect to one latent trait can make up for a lack in another (Junker 1999). In our model, we utilize the latent topics in the domain as a representation of the latent traits to be tested; due to the presence of overlap between these topics, we adopt a *compensatory* model, where it is assumed that knowledge of some topics can make up for a lack of knowledge in others.

In addition, we adopt some additional assumptions about learning which have been shown to be consistent with a broad-class of theories (Pirolli and Wilson 1998). We assume that knowledge content (and thus latent ability) grows monotonically with respect to each of these topics and in discrete quanta through interaction with the environment. Thus, we assume that a user's knowledge within a domain only increases; for simplicity, we ignore the effects of forgetting and of "throwing away" inconsistent bits of information.

3.3 Measurement model for the Nelson et al. (2009) SparTag.us study

Using the topic model, one can construct a set of semantic topics against which users and test items can be measured. In order to use these topics to measure and predict changes in states of topic knowledge as a function of learning, we need to integrate the semantic information given by the topic model with a version of the item response model (Pirolli and Wilson 1998) described above. We characterize the elements comprising the situation to be modeled below:

- A set of M individuals, indexed $i = 1, 2, \ldots, M$.
- A set of N test items on which individuals have been tested, indexed $j = 1, 2, \ldots, N$
- A set of T topics derived from the test items and documents read by the individuals, indexed $k = 1, 2, \ldots, T$.
- A set of $(M \times N)$ test item responses indexed Y_{ij} corresponding to the observed responses of each of the M individuals to each of the N test items.

Each response from an individual i to a test item j is coded binomially, as follows:

$$Y_{ij} = \begin{cases} 1, \text{ if participant } i \text{ answers test } j \text{ correctly} \\ 0, \text{ otherwise} \end{cases}$$

As is common in item response models for binomial data, we define the probability of a correct response using a logistic function.

$$\Pr\left(Y_{ij} = 1 | \tilde{\theta}_i, \tilde{w}_j, \rho_{ij}\right) = \frac{1}{1 + e^{-f\left(\tilde{\theta}_i, \tilde{w}_j, \rho_{ij}\right)}} \tag{7}$$

where the terms are defined as follows:

- $\tilde{\theta}_i$ represents the set of *ability* parameters for each individual i: some of these parameters will be estimates of latent knowledge or ability and some will capture learning ability.
- $\tilde{w}_j$ represents the distribution of topics which need to be accessed when answering a particular test item j.
- ρ_{ij} represents the relevance of reading performed by an individual i to the mix of topics involved in answering test item j (a measure of *precision-relevance* of reading to the test item).
- f is a linear function combining the parameters relevant to Y_{ij}.

This definition has the property that the log-odds formulation:

$$\log\left[\frac{\Pr\left(Y_{ij} = 1 | \tilde{\theta}_i, \tilde{w}_j, \rho_{ij}\right)}{\Pr\left(Y_{ij} = 0 | \tilde{\theta}_i, \tilde{w}_j, \rho_{ij}\right)}\right] = f\left(\tilde{\theta}_i, \tilde{w}_j, \rho_{ij}\right) \tag{8}$$

can be fit using generalized linear modeling techniques.

The ability parameters that we will estimate for each individual i will include

- *Latent Knowledge or Ability:* A set $\theta_{i1}, \ldots, \theta_{iT}$ of parameters which represent the strength of knowledge of an individual i with respect to each of the T topics.
- *Overall Learning Rate:* A parameter corresponding to the rate at which an individual learns material across the subject matter domain. Below, in our first model (Model 1), we will assume a single learning rate parameter θ_0 that is uniform across users (i.e. all users learn at the same rate). In our second model below (Model 2), we will use learning rate parameters θ_{i0} which are assumed to vary across individuals

(i.e. each individual learns at his or her own rate). Model 1 is more parsimonious than Model 2 (having fewer parameters), however, Model 2 has the benefit of yielding individualized estimates of learning, which is usually what is needed in knowledge tracing in ITSs.

In order to define the mix $\tilde{w}_j$ of topics involved in producing a correct answer to a question j, we define w_{jk} as the proportion of question j which corresponds to knowledge about topic k, where

$$\sum_{k=1}^{T} w_{jk} = 1$$

To define the amount of reading ρ_{ij} performed by an individual i that is relevant to a test item j, we first define a topic specific reading parameter r_{ik} to be the proportion of words read by the individual i which were assigned by the topic model to the topic k, where

$$\sum_{k=1}^{T} r_{ik} = 1$$

This reading parameter for each individual and topic can be estimated by aggregating the topic distributions as estimated by the topic model for each of the documents read by that individual. Because individuals may not read all of the words on a given web page, by estimating r_{ik} in this way, we make the implicit assumption that content pertaining to topics are uniformly distributed within a page. By looking at the overall distribution of reading that individuals performed with respect to each topic and the mix of topics tested by each item, we can define how the precision-relevance of the individuals' reading to a particular test item using the formula

$$\rho_{ij} = \begin{cases} 0, \text{ if item } j \text{ is on the pretest} \\ \sum_{k=1}^{T} r_{ik} w_{jk}, \text{ if item } j \text{ is on the posttest} \end{cases}$$

Thus, the parameters w_{jk} are determined using the topic model, and the parameters ρ_{jk} are determined by incorporating the reading data. We then estimate latent abilities with respect to topics and learning rates of individual users by using generalized linear model estimates of $\tilde{\theta}_i$.[1]

[1] We developed models that used *total* amount of reading per topic, rather than the *proportion* of reading per topic, but those models yielded results virtually identical to the models reported here. Specifically, the AIC goodness-of-fit metric for all models except for Model 1 are identical whether one uses "proportion read per topic" or "total read per topic". For Model 1, the AIC for the "proportion read" model is AIC = 853.72 which is marginally better than the "total read" AIC = 858.71. For Model 2, the shift from "proportion read per topic" to "total read per topic" changes the θ_{i0} learning rate estimates, although the estimates are highly correlated at $r = 0.92$. All other Model 2 parameter estimates are identical. Interestingly, the "total read" Model 2 θ_{i0} learning rate estimates are negatively correlated, with total words read, $r = -0.54$, which suggests a diminishing returns to reading about a specific topic.

4 Modeling latent topics in the social reading dataset

4.1 Document corpus

In order to produce the topic model, we drew content from a document collection consisting of the following:

- *Documents read during learning (Web pages):* Each participant's Web browsing session was logged, including URL's visited, content scrolled over, and words written. The text of each visited page was later captured and stored in the database. In total, the 18 participants visited $N = 1{,}759$ Web pages, with 1,146 distinct URL's, as some URL's were visited by multiple participants. Each participant viewed an average of 97.7 pages per session and a median of 99.5 pages per session.
- *Documents produced after learning (Essays):* Participants were informed prior to the learning task that they would be asked to complete two essay tasks afterwards. All participants had access to the Web and their notes (through either *SparTag.us* or other note-taking tools depending on condition) during the writing period. Both essays from each of the 18 participants were included, for a total of $N = 36$ documents.
- *Documents in the 'expert' notebook (Web pages):* The notebook of the simulated friend 'mjones' consisted of annotations comprising a tag cloud, a list of URL's, and a set of paragraphs, with the goal of crafting clear and succinct pointers to key content using social sources. The list of pages to be annotated were derived from 20 tags associated with the top 100 annotated URL's from a *delicious* query for the search term "enterprise mashup", which provided the target tag cloud. Each of these tags was then entered into a Google search, and a subset of the returned URL's chosen by an expert were stored in a notebook and then manually tagged using *SparTag.us* in order to provide the annotations. The resulting notebook thus simulated tags and annotations on $N = 22$ distinct Web pages, which were included in the document collection. As delicious tags are user-assigned, it should be noted that these tags may not have appeared in the documents themselves.
- *Documents associated with the test questions (Web pages)*: Because of the short length of the test items (each was a single sentence), modeling the mix of topics associated with each question required elaboration with additional relevant text. Thus, for each of the test items, we algorithmically identified ten Web pages containing data relevant to answering that question. Drawing on the term-expansion approach used in Bernstein et al. (2010), noun phrases were extracted for each statement (or an expert-substituted 'true' version for false statements) using the Stanford Part-of-Speech Tagger[2] and entered as a query into the Yahoo! BOSS Search Engine.[3] The top ten results for each of the 40 test items ($N = 400$) were captured and the text parsed for inclusion in the document collection.

[2] http://nlp.stanford.edu/software/tagger.html.

[3] http://developer.yahoo.com/search/boss/.

4.2 Inducing the topic model

For each document in the dataset, all text was passed through a stemmer and stopword filter built on the public domain NLTK toolkit.[4] To develop the topic model, we utilized a standard methodology which we describe below—for additional information, please cf. (Griffiths et al. 2007). The stemmed and filtered documents were then transformed into a word-token vector, which mapped each individual token to a distinct vocabulary word, and a document-token vector, which mapped each token to the document in which it originally occurred. The final word-token and document-token vectors contained 3,115,628 individual tokens, mapped to $D = 2{,}217$ documents and $W = 48{,}418$ distinct vocabulary words. LDA was performed using Gibbs Sampling via the MATLAB Topic Modeling Toolbox 1.3.2.[5] The sampling algorithm was run for 500 iterations with hyperparameters $\alpha = T/50$ and $\beta = 200/W$, values which have been chosen based upon exploration in previous studies involving topic modeling on similar document corpora (Griffiths et al. 2007).

A common method for choosing the number of topics is by calculating the perplexity, which is a standard measure for estimating the performance of a probabilistic model (c.f., Blei et al. 2003; c.f., Rosen-Zvi et al. 2004). Essentially, for a model trained on a set of training data, the perplexity is a measure of how well the model generalizes to a different set of test data. Specifically, if the model assigns a high likelihood to the data encountered in the test data, this will result in a lower perplexity value, which indicates greater generalization performance. While model perplexity decreases monotonically as T increases, we also seek to minimize T in order to avoid over-fitting. Thus, the optimal value of T is chosen where the graph "bends".

Using such a perplexity-based approach, we initially chose $T = 100$ as our 'optimal' number of topics. However, initial exploration revealed that the relative sparsity of our item response data required the use of a much smaller number of topics in our analysis. The response data for the M = 18 participants and $N = 40$ test questions yielded a total of 720 responses, which represent the main dependent variables to be predicted by the model. Fitting the data to our first model (Model 1) entails the estimation of $(M \text{ x } T) + 1$ parameters, while the second model requires even more, meaning that we had to limit the number of topics in our analysis to $T = 5$ in order to preserve degrees of freedom. However, given more data points to fit, there is no reason why this same approach would not work with a greater number of topics. While the topics generated using $T = 5$ were statistically distinguishable from one another, we do not present them in detail here, as the coarse nature of these topics makes them difficult to interpret semantically.

[4] http://nltk.org/.

[5] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm.

5 Applying the measurement framework and comparing model variations

5.1 Model 1: individual differences in background and common learning rates

In our first model, which we refer to as Model 1, we include a separate latent ability parameter for each individual on each topic and we assume a single learning rate parameter across all individuals. Thus, for any individual i, we define the linear function f described above as

$$f\left(\tilde{\theta}_i, \tilde{w}_j, \rho_{ij}\right) = \theta_0\rho_{ij} + \theta_{i1}w_{j1} + \cdots + \theta_{iT}w_{jT} \tag{9}$$

In other words, this model assumes that all users acquire topic knowledge at the same rate, θ_0, with no individual differences. Because of the limited set of response data, we used a topic model with just $T = 5$ topics. It should also be noted that through this paper we fit learner parameters as fixed effects. A more appropriate method, given sufficient data, would be to treat learner parameters as random coefficients to be estimated (cf., the multidimensional random coefficients multinomial logit model, Pirolli and Wilson 1998).

Figure 6 presents a plot of individual background knowledge profiles. These are estimated based on the overall fit of the model to the data. Participants (1…18) are plotted along the x-axis, topics (1…5) along the y-axis, and the strength of background knowledge is in the z-axis. Even with just 5 topics, it is clear that there is a large amount of variation in background knowledge with respect to the different topics.

Figure 7 presents the predicted performance (dashed line) on a hypothetical Post-Test item as a function of the precision, ρ_{ij} of relevant reading (x-axis) performed by a user. The learning rate estimated in the logistic regression of Model 1 was θ_0 = 3.36. One can observe how the probability of a correct response increases from about $\Pr(Y_{ij} = 1) = 0.5$ (which is the baseline guessing rate, up to $\Pr(Y_{ij} = 1) = 1.0$ as the relevance of the read material to the test question increases to a precision of $\rho = 1.0$. Circles plot the expected probability correct values of ρ_{ij} for the N= 360 posttest observations in the dataset.

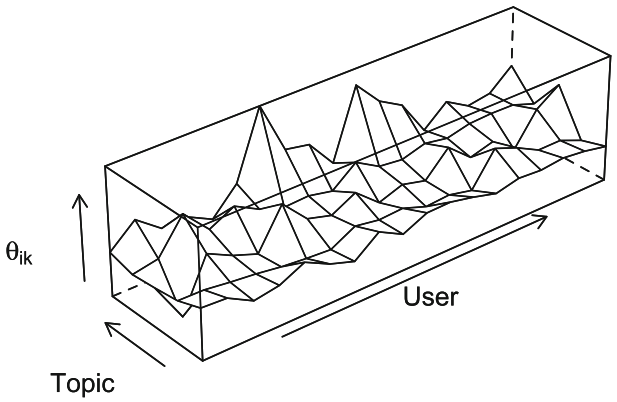

Fig. 6 Profile of background knowledge across users (x-axis) with respect to topics (y-axis). θ_{ik} is an estimate of topic knowledge strength produced by fitting to Model 1

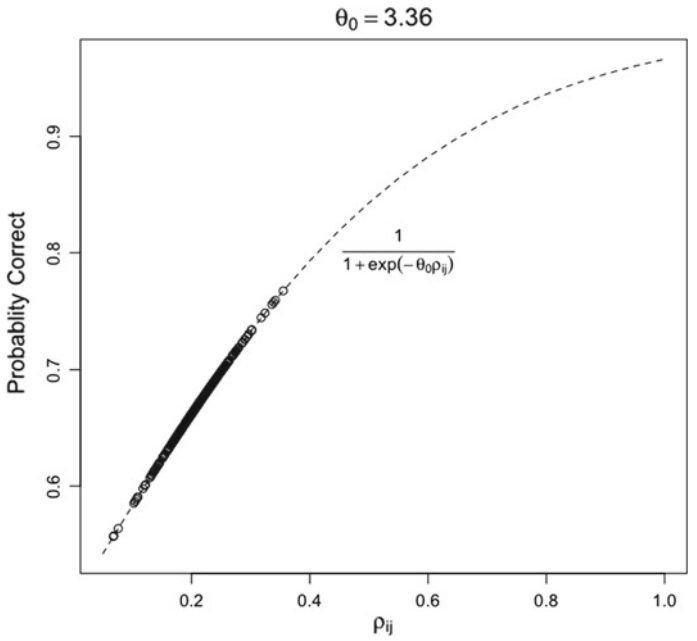

Fig. 7 The predicted performance on a hypothetical Post-Test item as a function of the precision of relevant reading (x-axis) performed by a user. The learning rate estimated in the logistic regression of Model 1 was $\Theta_0 = 3.36$. One can observe how the probability of a correct response increases from about $\Pr(Y_{ij} = 1) = 0.5$ (which is the baseline guessing rate, up to $\Pr(Y_{ij} = 1) = 1.0$ as the relevance of the read material to the test question increases to a precision of $\rho = 1.0$

Using the parameter estimates from Model 1 of user knowledge states, θ_{ik}, the estimated learning rate, θ_0, and the mix of topics and the observed users' reading relevant to each test item, we can check how well the predicted response rates fit the observed performance. Figure 8 plots pre-test scores predicted by Model 1 against the observed scores. Each point plots the aggregate score observed for each participant ($max = 20$) as a function of the the aggregate number of correct responses predicted for that participant. The predicted Model 1 response rates in Fig. 8 are based solely on the background knowledge estimates (Fig. 7) for individual users. The correlation between predicted and observed pretest scores is $r = 0.95$, $t(16) = 12.017$, $p \ll 0.00001$

Figure 9 presents a similar plot of the observed post-test scores as a function of the predicted scores. In this case, the predicted scores include not only the background knowledge, but also the effects of learning from reading. The correlation between predicted and observed posttest scores is $r = 0.91$, $t(16) = 8.544$, $p \ll .00001$.

Figure 10 presents observed learning gains as a function of the predicted learning gains, where the learning gains are calculated as described in Sect. 2.3.3. Each point in Fig. 10 corresponds to an individual user. The correlation between predicted and observed learning gains is $r = 0.51$, $t(16) = 2.3621$, $p = .03$.

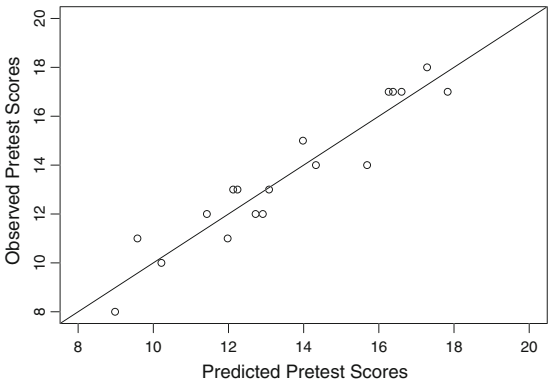

Fig. 8 *Scatter plot* of observed user pre-test scores against scores predicted by Model 1. *Line* (in this and subsequent figures) represents perfect prediction rather than the best-fitting line

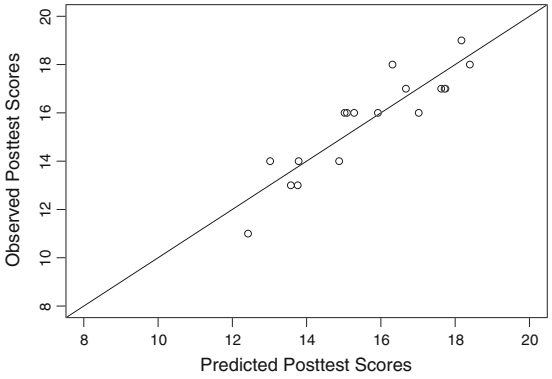

Fig. 9 *Scatter plot* of observed user post-test scores against scores predicted by Model 1

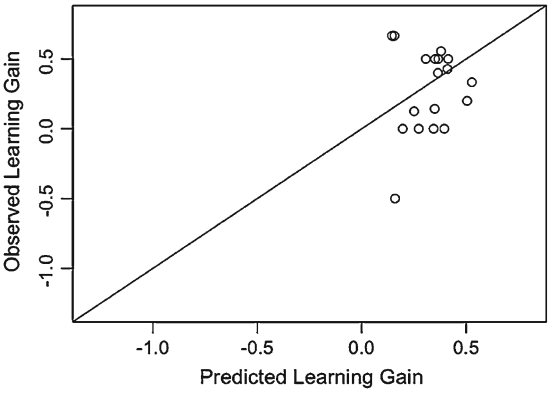

Fig. 10 *Scatter plot* of observed learning gains against Model 1 predicted learning gains

5.2 Model 2: individual differences in background plus differences in learning rates

Whereas Model 1 assumed that all users learned at the same rate from browser materials, the second model (Model 2) included a separate learning rate parameter for each individual user. In this model, f is specified as

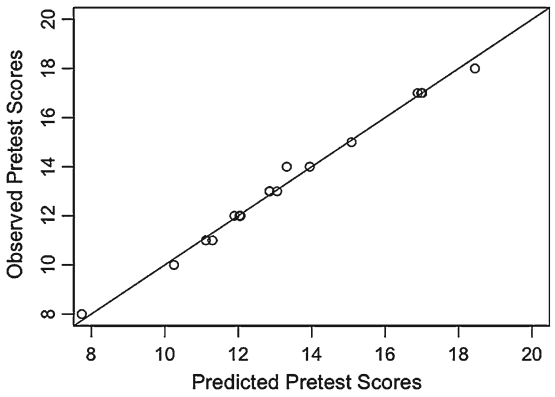

Fig. 11 *Scatter plot* of observed user pre-test scores against scores predicted by Model 2

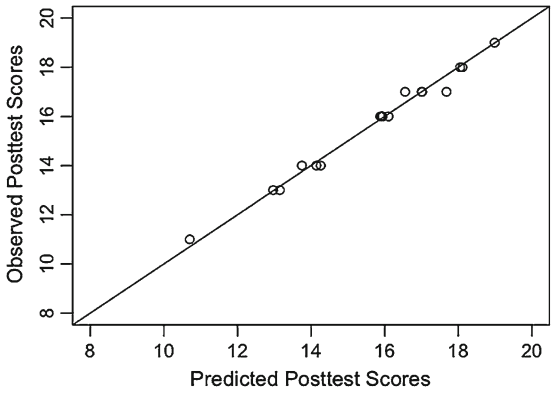

Fig. 12 *Scatter plot* of observed user post-test scores against scores predicted by Model 2

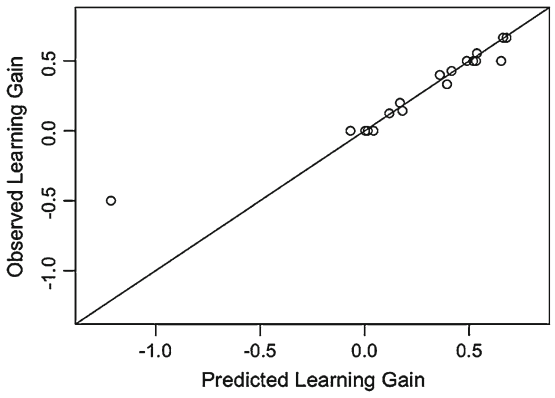

Fig. 13 *Scatter plot* of observed learning gains against learning gains predicted by Model 2

$$f\left(\tilde{\theta}_i, \tilde{w}_j, \rho_{ij}\right) = \theta_{i0}\rho_{ij} + \theta_{i1}w_{j1} + \cdots + \theta_{iT}w_{jT} \tag{10}$$

where the learning rate parameter θ_{i0} was free to vary across users i.

Figures 11 and 12 present scatter plots of the observed pretest and posttest scores against the Model 2 predictions for those scores. Figure 13 presents a scatter plot of

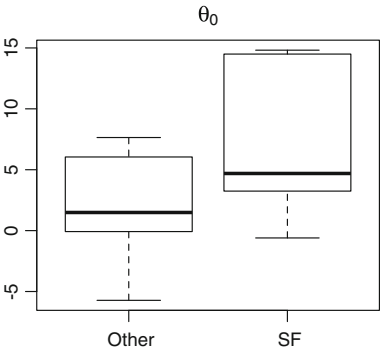

Fig. 14 *Box plot* of the learning rate parameters as estimated by Model 2. Learning rates were marginally greater for users with access to "expert" tags (SF group)

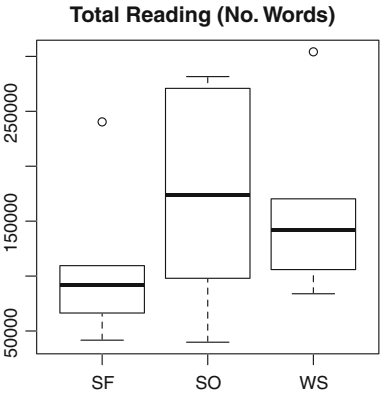

Fig. 15 *Box plot* of the total number of words in the documents browsed by users. Members of the SF group tended to read less than the groups without access to expert tags

the observed learning gains against those predicted by Model 2. A visual comparison of Fig. 13 for Model 2 to the analogous plot for Model 1 in Fig. 10 suggests much more accurate predictions by Model 2. The learning gains predicted by Model 2 have a high correlation with those observed (Fig. 11), $r = 0.95$, $t(16) = 12.4026$, $p \ll .00001$.

Model 2 provides additional insight into the increased learning efficiency of the SF group (with the 'expert' tags). Here, we note that instead of measuring efficiency in the amount of time required to learn a certain amount, we measure efficiency and learning rates in terms of the amount of text browsed.The learning rates (Fig. 14) for the SF group were generally higher than the other two groups [marginally significant, $F(1, 16) = 3.92$, $MSE = 23.19$, $p = .06$].

These differences in learning rates reflect some general observations about the reading performed by users in the different groups. Figure 15 shows that the SF group, with access to the "expert" tags, read less overall. Figure 16 shows that the SF group read materials that were more relevant to the post test items they encountered. In general, the data suggest the SF group showed a greater efficiency in reading higher precision materials as a function of browsing interaction.

Model 2 appears to provide more accurate predictions than Model 1, but it also uses more parameters to do so (at a loss of degrees-of-freedom of residuals). To compare the goodness-of-fit of the models, taking into account the complexity (degree-of-free-

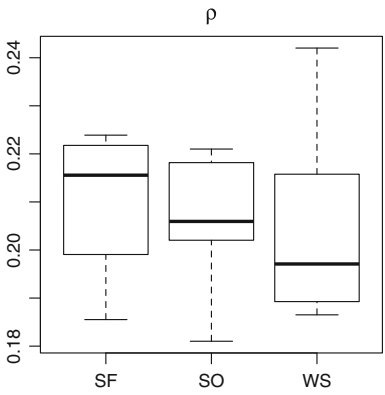

Fig. 16 *Box plot* of the precision of materials read by users with respect to their Post-Test questions. Members of the SF group tended to read more relevant material than the other groups

dom), we constructed the *analysis of deviance* presented in Table 2, which compares Models 1 and 2, along with several variants that we discuss below.

In an analysis of deviance, alternative nested models are ordered form fewest parameters to most $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$. For each model the associated residual deviance of the data from the model is calculated, $dev_1, dev_2, \ldots dev_n$(lower deviance means better fit). The degrees-of-freedom of the residuals of the models will decrease as the number of model parameters increases, $df_1 > df_2 > \cdots > df_n$. The differences in deviance (dev_n minus dev_{n-1}) will asymptotically be distributed as a χ^2 distribution with residual degrees of freedom df_n minus df_{n-1}. The models in the rows of Table 2 are arranged from top to bottom in order of increasing complexity (from fewest parameters to most). Table 2 reports the residual deviance and residual degrees-of-freedom for each model. Also, each row reports the change in $d.f.$ and change in deviance for a model in comparison to the model in the row above (columns 5 and 6). The penultimate column reports the significance of the improvement in model fit over the model in the prior row as assessed by χ^2 test.

The analysis of deviance comparison of Model 1 and Model 2 is not significant (shown in the last row and penultimate column of Table 2), suggesting that one can gain parsimony while retaining predictive power by assuming a uniform learning rate among subjects. However, Model 2 is still useful as it provides additional insight into why the participants in the SF group (those with access to 'expert' annotations) learned more efficiently than those in other groups. as shown in Fig. 12. In addition, in practice, it is useful to have individual knowledge and learning parameters to drive the instruction in an e-learning system (e.g., Corbett et al. 1995).

5.3 Models 3–5: no topic model

To assess the contribution of the LDA topic modeling, we compared Models 1 and 2 to three variations of the model which did not involve topic information.

– *Model 3:* This model assumes that individuals differ only with respect to a single background ability parameter (i.e. only a single background ability parameter θ_i was estimated for each user). In other words, this background ability parameter

Table 2 Comparison of models by analysis of deviance

Model	Terms	Estimated parameters	Residual d.f.	Residual deviance	df	Change in deviance	p	AIC (Cp)
Model 3	θ_i	18	702	793.61				829.61
Model 4	$\theta_i + \Delta$	19	701	782.70	1	10.909	0.0009568*	820.70
Model 5	$\theta_i + \Delta_i$	36	684	774.44	17	8.263	0.9605916	846.44
Model 1	$\theta_0 \rho_{ij} + \theta_{i1} w_{j1} + \cdots + \theta_{iT} w_{jT}$	91	629	671.82	55	10.719	0.0001014*	853.72
Model 2	$\theta_{0i} \rho_{ij} + \theta_{i1} w_{j1} + \cdots + \theta_{iT} w_{jT}$	108	612	656.60	17	15.122	0.5867008	872.60

Models 1 and 2, based on Topic Models, yield significantly better fits than Models 3–5
AIC Akaike Information Criterion, *Cp* Mallows' Cp statistic. AIC and Cp give the same values for these models
Notes: * Significant improvement in model fit compared to model in the row above

simply captures the overall propensity of an individual to know about "enterprise 2.0 mashups". In this model, we do not estimate any parameters corresponding to the reading performed or its relevance to the test item. In other words, no learning is assumed. for each individual, this model predicts a uniform likelihood of a correct response across all pre-test and post-test items.

– *Model 4:* This model assumes that individuals differ with respect to background ability, θ_i, and assumes a single parameter Δ, which is added to capture overall learning effects as a result of reading between pre-test and post-test. Unlike the learning rate parameters in Models 1 and 2, the Δ learning effect is not sensitive the particular reading histories of individuals. This learning parameter is assumed to be uniform across all subjects. . In other words, Δ simply captures the overall "treatment effect" of being exposed to the availability of reading materials.
– *Model 5:* This model assumes individual background ability parameters, θ_i, and learning parameters, Δ_i, which represent individual differences in the amount of information learned between pre-test and post-test. Again, this learning parameter is not associated with the particular reading histories of individuals. Each individual has a different prior background ability, and a different propensity to learn. However, the model is not sensitive to anything specific about what is read.

As shown in Table 2, there is a significant improvement in model fit as one goes from Model 3, which assumes that individuals differ only with respect to a single background ability parameter θ_i, to Model 4 by including a uniform pre-test to post-test learning effect Δ. The analysis of deviance indicated that there was no further significant improvement made by assuming individual pre-test to post-test learning effects Δ_i (i.e. going from Model 4 to Model 5).

5.4 Summary of model comparisons

Overall, good fits are achieved by a model (Model 1) that includes estimates of individual user background knowledge of topics and a learning rate parameter that is uniform across users. Increasing the complexity of the model to include individual differences in learning rates (Model 2) provides insights concerning the efficiency of learning for subjects provided with expert tags. The models based on topic models (Models 1–2) fit significantly better than those not involving topic models (Models 3–5). It is worthwhile noting the AIC (Akaike Information Criterion) scores (Table 2) would suggest that Model 4 provides a parsimonious best fit to the data, but that model simply captures individual differences in background ability and a single "treatment" effect of exposure to the Web. Such a model would not be very useful in providing individualized user-model adapted instruction.

6 General discussion

Assessing users's knowledge states is a critical component in developing web-based technologies to support automated e-learning systems. In this paper, we have described a framework for automatically inducing semantic topics represented in a domain, for

assessing knowledge states across multiple topics, and for assessing learning from browsing histories. Using data from an existing study of learning with a social tagging system (Nelson et al. 2009), we have demonstrated that topic models can form the basis of knowledge tracing.

As the item response data available were limited and we were forced to work with a small number of topics, future work includes applying this approach to larger data sets and in other e-learning systems in order to ensure its extensibility. A related area for future investigation is understanding why our model still yielded such high predictive accuracy given the small number of topics and whether a fine- or coarse-grained model actually yields better results. In addition, substantial work is required to develop this into a practical e-learning system. Such work includes determining a method of automatically identifying expert taggers (Noll et al. 2009)and automated methods for providing feedback (Foltz et al. 1999; Kakkonen et al. 2005).

An additional challenge related to assessment is in the automatic generation of test items for arbitrary knowledge domains. It is possible that this challenge could be addressed by leveraging user data from other sources; recent work has demonstrated that seeding user models using data from other social web systems for the purposes of personalization may be a promising approach (Abel et al. 2012; Shapira et al. 2012). As these challenges are surmounted, there is great potential for e-learning systems to tap the vast amounts of socially data available in systems similar to delicious in order to provide guidance for novice learners.

By applying this measurement framework to data from the social reading study, we gained additional insight into the learning effects experienced by users of the *SparTag.us* social annotation system. Our model was able to accurately predict not only the individual learning gains, but also helped to explain the learning differences between the groups in the study. In particular, our application of the measurement framework suggests that the efficacy of the *SparTag.us* system appears to lie in improving the precision of material that users read by providing access to expert 'signposts' to important content.

We utilized an analysis of deviance approach to compare model variations in order to find a version of the model which maximized both predictive power and degrees of freedom. We found that Model 1, which incorporated topical information but assumed that users all learned at the same rate, was able to predict post-test performance with a high degree of accuracy. Assuming individual differences in learning rates (Model 2), however, provided additional insight into the greater efficiency of learning for subjects provided with expert annotations. The results of this model-based analysis suggest that these types of learning improvements can be detected and predicted by this combination of topic modeling and psychometric measurement.

6.1 Discovering knowledge in social systems

Various approaches have been used to algorithmically discover expertise in social systems. In the case of social tagging applications, specifically, many of these approaches (Hotho et al. 2006; John and Seligmann 2006; Noll et al. 2009) have leveraged the graph structure of these systems, utilizing users, resources, and tags as nodes and the

connections between them (a resource receiving a certain tag, for example) as edges. Other approaches, such as Budura et al. (2009) have utilized probabilistic language models to represent expertise based on the tags used by a particular user in the system.

Probabilistic language models have also been used to identify expert people or resources in other social information systems, in particular those in which users are generating documents rather than simply annotating them. Some of these methods (e.g., Balog et al. 2006; Petkova and Croft 2006) have adapted generative probabilistic language modeling techniques in order to model potential experts in an enterprise using documents associated with them and matched them with a particular search query. Mimno and McCallum (2007), for example, applied a variation of the topic model to academic papers in order to extract an author's areas of expertise and match him or her (as a reviewer) with a particular paper.

Different social systems and knowledge sharing goals may call for different combinations of these approaches. In this volume, for example, Kim and El Saddik (2012) present a review of how several graph-based, language-based, and other techniques can be applied to the same goal of recommending communities of interest based on user tags. Given a task setting an the appropriate combination of approaches for the identification of expert users or documents, such resources could provide the inputs into an automated e-learning system built on the framework described in this paper. By modeling users through flexible latent topics rather than rigid semantic networks, we can easily identify new resources, characterize their gist, and decide which ones will best fill topical gaps in knowledge for learners.

6.2 Applications of this work to intelligent tutoring systems

In this paper, we have described a framework for inducing the semantic topics in a domain and using them to assess the knowledge states of users learning in this domain while browsing Web-based resources. With the exception of the expert-formulated test items, our measurement framework was applied in a situation in which the domain corpus was generated entirely using resources available in social information systems. We have shown this framework to be capable of accurately predicting learning gains and post-test performance for subjects in the *SparTag.us* social reading study and that the inclusion of topic information significantly increases the predictive power of these models.

6.3 Limitations

Topic models, though useful, are limited in several ways in their ability to represent knowledge and expertise. Many ITSs focus on the representation of cognitive skill in domains such as algebra (Koedinger et al. 1997) or programming (Corbett et al. 1995). Such cognitive skills are procedural knowledge (Anderson et al. 2004) of *how* to do things, such as moving symbols in an equation or writing syntactically correct code that iterates through a list. Topic models, however, have been used in psychological models of *declarative* knowledge such as subject matter facts and concepts. In theories such as ACT-R (Anderson et al. 2004), procedural knowledge is represented in one

form (production rules) and declarative knowledge in another (semantic networks of chunks). For semantically rich domains that require mastery of large amounts of both declarative knowledge and procedural knowledge, topic models might focus on the declarative part, and complement other representations that focus on the procedural part. A related limitation is that this approach does not account for information which must be learned in a particular sequence, as the model treats all topics identically.

6.4 Future work

SparTag.us and the Nelson et al. (2009) study were not originally motivated by an eventual goal of developing an e-learning system. However, the system actually provides behavioral and content traces that are sufficient to do knowledge tracing. Two of the significant things done by hand in the Nelson et al. (2009) study include the identification of "expert" tags and documents, and the construction of test items. As we have noted throughout, there are a number of techniques (e.g., Noll et al. (2009)) emerging for expert-finding in social tagging systems. Test items themselves may be difficult to construct, but Foltz et al. (1999) have demonstrated that free-form essays can be automatically graded using Latent Semantic Analysis, and a similar scoring function might be built on top of LDA. It remains to be demonstrated that the kind of Web-based e-learning system sketched here could achieve the one- to two-sigma gains achieved by some ITSs, but what we are proposing has a significant "economic" advantage over current ITSs: whereas current ITSs require substantial knowledge engineering, the Web-based model proposed here involves automatic induction of expert models and user models based on crowd wisdom in social tags.

Acknowledgments This research was supported by a contract from the Office of Naval Research: Contract No. N00014-08-C-0029 to Peter Pirolli. We would also like to acknowledge Gregorio Convertino, Lichan Hong, Les Nelson, and the rest of the Augmented Social Cognition Group at PARC for their contributions to this research.

References

Abel, F., Baldoni, M., Baroglio, C., Henze, N., Krause, D., Patti, V.: Context-based ranking in folksonomies. Paper presented at the proceedings of the 20th ACM conference on hypertext and hypermedia (HT '09), pp. 209–218 (2009)

Abel, F., Herder, E., Houben, G.-J., Henze, N., Krause, D.: Cross-system User Modeling and Personalization on the Social Web. User Model. User Adapt. Interact. (2012). doi:10.1007/s11257-012-9131-2

Aleven, V., McLaren, B.M., Sewall, J., Koedinger, K.R.: A new paradigm for intelligent tutoring systems: example-tracing tutors. J. Artif. Intell. Educ, **19**(2), 105–154 (2009)

Allen, I.E., Seaman, J.: Learning on demand: online education in the United States, 2009: Sloan Consortium (2010)

Anderson, J.R.: Learning to program in LISP. Cogn. Sci. **8**, 87–129 (1984)

Anderson, J.R.: The Adaptive Character of Thought. Lawrence Erlbaum Associates, Hillsdale (1990)

Anderson, J.R.: Rules of the Mind. Lawrence Erlbaum Associates, Hillsdale (1993)

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of mind. Psychol Rev **11**(4), 1036–1060 (2004)

Anderson, J.R., Boyle, C.F., Corbett, A., Lewis, M.W.: Cognitive modelling and intelligent tutoring. Artif. Intell. **42**, 7–49 (1990)

Balog, K., Azzopardi, L., de Rijke, M.: Formal models for expert finding in enterprise corpora. Paper presented at the 29th annual ACM SIGIR conference on research and development in information retrieval (SIGIR '06) (2006)
Bernstein, M., Suh, B., Hong, L., Chen, J., Kairam, S., Chi, E.H.: Interactive topic-based browsing of social status streams. Paper presented at the proceedings of the 23rd symposium on user interface software and technology (UIST '10) (2010)
Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
Bloom, B.: The 2 sigma problem: the search for methods of instruction as effective as one-to-one tutoring. Educ. Res. **13**(6), 4–16 (1984)
Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. **30**(1–7), 107–117 (1998)
Brusilovsky, P., Peylo, C.: Adaptive and intelligent web-based educational systems. Int. J. Artif. Intell. Educ. **13**, 156–169 (2003)
Budura, A., Bourges-Waldegg, D., Riordan, J.: Deriving expertise profiles from tags. Paper presented at the international conference on computational science and engineering (CSE '09) (2009)
Campbell, C.S., Maglio, P.P., Cozzi, A., Dom, B.: Expertise identification using email communications. Paper presented at the proceedings of the 2003 ACM CIKM international conference on information and knowledge management (CIKM '03) (2003)
Carbonell, J.R.: AI in CAI: an artificial-intelligence approach to computer-assisted instruction. IEEE Trans. Man Mach. Syst. **11**(4), 190–202 (1970)
Champ, H.: (2009, August 13, 2010). 4,000,000,000. http://blog.flickr.net/en/2009/10/12/4000000000
Conati, C., Gertner, A., Vanlehn, K.: Using Bayesian networks to manage uncertainty in student modeling. User Model. User Adapt. Interact. **12**(4), 371–417 (2002)
Corbett, A.T.: Cognitive computer tutors: solving the two-sigma problem. Paper presented at the user modeling 2001: 8th international conference, Berlin (2001)
Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. User Model. User Adapt. Interact. **4**(4), 253–278 (1995)
Corbett, A. T., Anderson, J. R., O'Brien, A. T.: Student modeling in the ACT Programming Tutor. In: Nichols P. D., Chipman S. F., Brennan R. L. (eds.) Cognitively diagnostic assessment, pp. 19–41. Lawrence Erlbaum Associates, Hillsdale, NJ (1995)
Delicious.: (2008, August 13, 2010). Oh Happy Day. http://blog.delicious.com/blog/2008/07/oh-happy-day.html
Foltz, P.W., Laham, D., Landauer, T.K.: Automated essay scoring: application to educational technology. Paper presented at the world conference on education, multimedia, hypermedia, and telecommunications, Seattle, WA (1999)
Fox, S., Fallows, D.: Internet health resources. Retrieved December, 2003, from http://www.pewinternet.org/reports/pdfs/PIP_Health_Report_July_2003.pdf (2003, August)
Fox, S., Jones, S.: The social life of health information. Pew Internet & American Life Project. Retrieved June 27, 2009, from http://www.pewinternet.org/Reports/2009/8-The-Social-Life-of-Health-Information.aspx (2009, June 11)
Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T.: The vocabulary problem in human-system communication. Commun. ACM **30**, 964–971 (1987)
Gayo-Avello, D., Brenes D.J.: Overcoming spammers in twitter—-a tale of five algorithms. Paper presented at the Congreso Español de Recuperación de Información (CERI 2010) (2010)
Gena, C., Cena, F., Vernero, F., Grillo, P.: The evaluation of a social adaptive web site for cultural events. User Model. User Adapt. Interact. (2012). doi:10.1007/s11257-012-9129-9
Golder, S.A., Huberman, B.A.: The structure of collaborative tagging systems. J. Inf. Sci. **32**, 198–208 (2006)
Griffiths, T.L., Steyvers, M., Tenenbaum, J.B.: Topics in semantic representation. Psychol. Rev. **114**(2), 211–244 (2007)
Han, S.-G., Lee, S.-G., Jo, G.-S.: Case-based tutoring systems for procedural problem solving on the www. Expert Syst. Appl. Int. J. **29**(3), 573–582 (2005)
Hartley, J.R., Sleeman, D.H.: Towards more intelligent teaching systems. Int. J. Man Mach. Stud. **5**(2), 215–236 (1973)
Hofman, T.: Probabilistic latent semantic indexing. Paper presented at the twenty-second international SIGIR conference on research and development in information retrieval (SIGIR-99) (1999)

Hong, L., Chi, E. H., Budiu, R., Pirolli, P., Nelson, L.: SparTag.us: a low cost tagging system for foraging of web content. Paper presented at the working conference on advanced visual interfaces (AVI '08) (2008)

Horrigan, J.: The Internet as a resource for news and information about science. Pew Internet & American Life Project. Retrieved June 27, 2009, from http://www.pewinternet.org/Reports/2006/The-Internet-as-a-Resource-for-News-and-Information-about-Science.aspx (2006, November 20)

Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: search and ranking. Paper presented at the 3rd European semantic web conference (ESWC '06) (2006)

John, A., Seligmann, D.: Collaborative tagging and expertise in the enterprise. Paper presented at the collaborative web tagging workshop in conjunction with 15th international conference on world wide web (WWW '06) (2006)

Junker, B.: Some statistical models and computational methods that may be useful for cognitively-relevant assessment. National Research Council (1999)

Kakkonen, T., Myller, N., Timonen, J., Sutinen, E.: Automatic essay grading with probabilistic latent semantic analysis. Paper presented at the proceedings of the second workshop on building educational applications Using NLP (2005)

Kammerer, Y., Nairn, R., Pirolli, P., Chi, E. H.: Signpost from the masses: learning effects in an exploratory social tag search browser. Paper presented at the 27th conference on human factors in computing systems (CHI '09) (2009)

Kim, H.-N., El Saddik, A.: Exploring social tagging for personalized community recommendations. User Model. User Adapt. Interact. (2012). doi:10.1007/s11257-012-9130-3

Kleinberg, J.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)

Kline, T.J.B.: Psychological Testing: A Practical Approach to Design and Evaluation. Sage Publications, Thousand Oaks (2005)

Koedinger, K.R., Anderson, J.R., Hadley, W.H., Mark, M.A.: Intelligent tutoring goes to the big city. Int. J. Artif. Intell. Educ. **8**, 30–43 (1997)

Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. **104**, 211–240 (1997)

Lenhart, A.: Adults and social network websites. Pew Internet & American Life Project. Retrieved June 27, 2009, from http://www.pewinternet.org/Reports/2009/Adults-and-Social-Network-Websites.aspx (2009, January 14)

Loizou, S. K., Dimitrova, V.: Adaptive notifications to support knowledge sharing in close-knit virtual communities. User Model. User Adapt. Interact. (2012). doi:10.1007/s11257-012-9127-y

Mimno, D., McCallum, A.: Expertise modeling for matching papers with reviewers. Paper presented at the 13th international conference on knowledge discovery and data mining (KDD '07) (2007)

Murray, T.: Authoring intelligence tutoring systems: an analysis of the state of the art. Int. J. Artif. Intell. Educ. **10**, 98–129 (1999)

Murray, T: An overview of intelligent tutoring system authoring tools: updated analysis of the state of the art. In: Murray, T., Blessing, S., Aisworth, S. (eds.) Authoring Tools for Advanced Technology Environments, pp. 493–546. Kluwer, Dordrecht (2003)

Nelson, L., Held, C., Pirolli, P., Hong, L., Schiano, D., Chi, E. H.: With a little help from my friends: examining the impact of social annotations in sensemaking tasks. Paper presented at the proceedings of the 27th international conference on human factors in computing systems (2009)

Noll, M. G., Au Yeung, C.-M., Gibbins, N., Meinel, C., Shadbolt, N.: Telling experts from spammers: expertise ranking in folksonomies. Paper presented at the 32nd annual ACM SIGIR conference on research and development in information retrieval (SIGIR '09) (2009)

Noll, M. G., Yeung, C.-m. A., Gibbins, N., Meinel, C., Shadbolt, N.: Telling experts from spammers: expertise ranking in folksonomies. Paper presented at the proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval (2009)

PACT.: Pittsburgh advanced cognitive tutor center home. Retrieved August 13, 2010, from http://pact.cs.cmu.edu (2005)

Petkova, D., Croft, B.W.: Hierarchical language models for expert finding in enterprise corpora. Paper presented at the 18th IEEE conference on tools with artificial intelligence (ICTAI '06) (2006)

Pirolli, P., Wilson, M.: A theory of the measurement of knowledge content, access, and learning. Psychol. Rev. **105**, 58–82 (1998)

Robu, V., Halpin, H., Shepherd, H.: Emergence of consensus and shared vocabularies in collaborative tagging systems. ACM Trans. Web **3**(4), 1–34 (2009)

Rosen-Zvi, M., Griffiths, T.L., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. Paper presented at the 20th conference on uncertainty in artificial intelligence (2004)
Shapira, B., Rokach, L., Freilichman, S.: Utilizing facebook single and cross domain data for recommendation systems. User Model. User Adapt. Interact. (2012). doi:10.1007/s11257-012-9128-x
Shute V.J., Psotka J.: Intelligent tutoring systems: past, present, and future. In D.H. Jonassen (ed.) Handbook of Research on Educational Communications and Technology, pp. 570–600. Simon & Schuster Macmillan, New York (1996)
Steyvers M., Griffiths, T.L., Dennis, S.: Probabilistic inference in human semantic memory. TRENDS Cogn. Sci. **10**(7), 327–334 (2006)
Weng, J., Lim, E.-P., Jiang, J., He, Q.: TwitterRank: finding topic-sensitive influential twitterers. Paper presented at the proceedings of the 3rd ACM international conference on web search and data mining (WSDM '10) (2010)
Yin, Z., Li, R., Mei, Q., Han, J.: Exploring social tagging graph for web object classification. Paper presented at the 15th ACM SIGKDD conference on knowledge discovery and data mining (KDD '09) (2009)
YouTube. YouTube Statistics. Retrieved January 16, 2012, from http://www.youtube.com/t/press_statistics
Zhang, J., Ackerman, M.S.: Searching for expertise in social networks: a simulation of potential strategies. Paper presented at the 2005 international ACM SIGGROUP conference on supporting group work (Group '05) (2005)
Zhang, J., Ackerman, M. S., Adamic, L.: Expertise networks in online communities: structure and algorithms. Paper presented at the 16th international world wide web conference (WWW '07) (2007)

Author Biographies

Peter Pirolli is a Research Fellow in the Augmented Social Cognition Area at the PARC, where he has been pursuing studies of human information interaction since 1991. Prior to joining PARC, he was an Associate Professor in the School of Education at UC Berkeley. Pirolli received his doctorate in cognitive psychology from Carnegie Mellon University in 1985. He is an elected Fellow of the American Association for the Advancement of Science, the Association for Psychological Science, the American Psychological Association, the National Academy of Education, and the Association for Computing Machinery Computer-Human Interaction Academy.

Sanjay Kairam is a graduate student in the Computer Science Department at Stanford University. Advised by Jeffrey Heer, his research focuses on visual and statistical analysis of human behavior in online networks and communities. The work in this volume was conducted while he was serving as a research assistant with the Augmented Social Cognition group at PARC. Sanjay received his B.S. in Mathematics and M.A. in Philosophy in 2006, also from Stanford, along with a minor in the interdisciplinary Symbolic Systems program.