Smarter learning content management using the Learning Content Hub

D. Contractor
S. Negi
K. Popat
S. Ikbal
B. Prasad
S. Vedula
S. Kakaraparthy
B. Sengupta
V. Kumar

The education sector is experiencing an unprecedented shift in how students learn and progress through their education. With the large amounts of digital learning content available, teachers are increasingly turning to sources such as online tutorials and eBooks for their teaching needs. Schools often make use of a Learning Content Management System (LCMS) to store learning material, which is catalogued based on the curriculum followed in the school. However, the amount of content indexed in the LCMS is often limited by the ability to manually label and catalogue content. In this paper, we describe our LCMS called the Learning Content Hub (LCH), which not only offers the features of a traditional LCMS including document search and retrieval, document security, user role management, etc., but also provides the ability to automatically analyze and label documents. LCH provides a framework for easily extending the analytics support and exposes application programming interfaces that can be used to build custom education applications for which the content needs can be met using the LCH. We discuss experimental results of the analyzers in our system as well as our experience of deploying this system in a U.S. school district.

Introduction

With less expensive and easier access to technology, the education industry has experienced a paradigm shift in the way learning content is created, organized, and distributed. For instance, today many K-12 schools in the United States use a Learning Content Management System (LCMS) [1] that has documents tagged based on the learning standard followed in their course curriculum. Academic learning standards [2] specify, through a series of discrete instructions, the skills that students should acquire as their learning progresses, and learning content is often retrieved using these instruction labels. "*Solve problems using concepts in graph theory, including directed and undirected graphs, the Handshaking Theorem, isomorphism, paths, connectedness, and Euler and Hamilton Paths*" is an example of a learning standard instruction from the High School Mathematics of the Academic Knowledge and Skills (AKS) learning standard [3]. Increasing use of learning management systems (LMSs) by schools and universities, coupled with rapid growth in learning content, necessitates that current LCMSs be more efficient in how they organize and deliver content.

The choice of content available to teachers and students for their teaching and studying needs is strictly limited to the content that has been aligned with the curriculum and vetted by the school district. With newer content being made available from publishers, teachers and students may want to refer to additional content for their teaching or learning needs. This content, however, may not be aligned to the curriculum being followed by the institution, thus making it difficult to be used directly in the teaching process. Manually identifying the appropriate learning standard instruction for each learning content item is time consuming and not scalable. Learning standards contain hundreds of topics per subject, and the standards themselves are subject to periodic revisions, making manual tagging a very costly and infeasible proposition.

Digital Object Identifier: 10.1147/JRD.2015.2455691

0018-8646/15 © 2015 IBM

We propose a novel learning content analytics component, which is part of the larger content management system described in this paper, which automatically tags content with the appropriate learning standard instructions. The automatic tagging feature makes our learning content management system unique in comparison to many of its counterparts. For example, in Ochoa et al. [4], metadata generated for each piece of learning content does not include learning standard instructions. Apart from supporting efficient content management, the content analytics components also makes our content management system an interesting candidate for competency-based learning [5–7] and personalized learning [5, 8], which require dynamic and need-based delivery of learning content. Aside from content analytics, the proposed learning content management system (referred to as Learning Content Hub, or LCH) provides native document access security, user role management, document workflows, search interfaces, APIs (application programming interfaces), etc. The LCH includes well designed user interfaces that facilitate document ingestion and retrieval. These user interfaces can be easily customizable based on user roles and needs.

In the next section, we outline the design principles that guided the development of LCH. In the section "System design and architecture of Learning Content Hub," we describe the system architecture and the content analytics components that are part of LCH. In the section on evaluation, we present an experimental evaluation of our content analytics component and share our experiences of deploying LCH in a U.S. school district. Finally, we conclude our paper and discuss future work.

Design principles

The following six design principles guided the LCH architecture. First, it was critical that adding new learning content or resource into LCH should be very easy for instructional staff, whether that content came from a publisher, from open educational resources, or from the direct authoring by teachers. Our primary criteria for ease of adding content concerned how little information would need to be explicitly stated at the time the content is registered with the catalog. We also want to support learning content sets, such as IMS (originally referring to Instructional Management Systems) specification content, often created by publishers [9].

As part of a second design principle, a software-assisted approach was needed to determine which learning standard topics are taught in each piece of cataloged content. We did not want to create a custom-developed hand-crafted approach to a single academic learning standard, because we anticipate numerous learning standards being adopted over time in many parts of the world. In order to allow automated labeling of content with learning standards, we needed to develop algorithms that could be executed within the LCMS system without requiring significant customization or redevelopment.

Third, the system should be able to allow deployment of custom document analyzers by the system users in the school district. In LCH, we support any analyzer that follows the Apache Unstructured Information Management Architecture (UIMA) standard [10]. For example, a user can develop or acquire video transcription engine to produce a transcript that will flow into our analyzer, and the output of the analytics will be stored as metadata in LCH.

Fourth, the LCMS system we developed had to consider the hierarchical nature while storing metadata as well as support one to many document-metadata relationships, since the learning standards tend to be hierarchical (topics and courses for different grades). We also wanted to be able to support regular document operations of addition, update, deletion, and retrieval, along with version control. In addition, the metadata and attributes for documents should be easily defined and edited, as LCH supports an open-ended set of UIMA analyzers.

Fifth, the system should be able to support the content servicing needs of any custom application. It is imperative that the system be customizable and extendable, and we expose data management and search APIs to address this requirement.

Lastly, our LCMS system provides traditional LCMS features such as keyword-based free text search, metadata based filtering, document security and user role management as well as content authoring or review workflows. The next section describes the system architecture and the implementation details that adhere to the design requirements summarized above.

System design and architecture of the LCH

The LCH consists of the following main components: (1) Data Ingestion, (2) User Role Management, (3) Document Lifecycle or Workflow Management, (4) Content and Metadata Storage, (5) Content Analytics, and (6) Search.

In a typical flow, content is ingested using the *Data Ingestion* component. Later, the content (and associated metadata) is stored and managed by the *Content and Metadata Storage* component. The ingested content is then passed through a series of annotators that extract relevant metadata from the content. This is handled in the *Content Analytics* layer. Overall user role management, security, and document workflow are orchestrated by the *User Role Management* and *Document Lifecycle or Workflow Management* system. Each of these components is described in the following subsections.

Data ingestion

The LCH acts as a secure repository for hosting and managing learning content. The LCH handles all aspects of the learning content life cycle. It provides a rich set of APIs to be used by content publishers or other vendors to either develop new applications or easily integrate their existing applications with LCH.

The LCH also provides crawlers that can be configured to crawl existing learning repositories. In addition, the LCH has a special crawler built-in to crawl IMS 1.2 "SCORM (Sharable Content Object Reference Model)" packages (a SCORM package is defined as a collection of content items, practice items, and assessment items that are combined based on a single learning objective) from host LMSs. This crawling capability will be extended to include other popular content standards such as Common Cartridge in the future. Once ingested, the crawled content is indexed and passed through a series of content analyzers for metadata enrichment and classification. Currently, the LCH does not have any content authoring and modification capabilities.

Customized ways of uploading content can also be configured on the LCH using active Web forms, referred to as *entry templates*. Entry templates in the LCH enable users to control how learning content is managed in the repository by defining the location, classification, property values and security. It saves time by providing the default information to streamline the process of adding new learning content. When using the entry template to add new content to the LCH repository, a user interacts with an intelligent user-interface, which decreases the chance of invalid data entry by limiting the number of steps required and providing a more controlled content upload process. The LCH also supports version control which allows users to access the most recent and all previous versions of a given document. An extensive audit trail of content changes is also maintained by the LCH.

Use role management

The LCH provides centralized access control over the learning content. Typical roles—such as curriculum coordinators, teachers, and students—can be easily modeled and managed in the LCH. Multiple security policies can be defined and assigned to these roles in the system to have better control and management of the learning content. Individual users can also create private workspaces to store private content or create user-groups with a defined set of users or roles for content sharing and review purposes.

Document lifecycle or workflows

Like any other content, the document lifecycle for learning content involves the following stages: creation, modification, review, and publish. Teachers, for example, might create content for a course, i.e., learning content that they would be using to teach a course. The learning content developed by teachers evolves over time. The LCH allows teachers to create private workspaces where the content draft can be stored privately and can be modified at their convenience. Once a teacher of a particular course is satisfied and ready with the learning content, it is available for sharing to the students enrolled to that course. However, some institutions might recommend a review of the learning content for quality checks and other copyrights compliance before publishing them externally. Thus, the LCH system also includes a workflow that can automate the review process. The teacher can submit the learning content to reviewers, e.g., a curriculum supervisor. Once the curriculum supervisor approves the quality of a learning content, it becomes available to be analyzed by the LCH, and it also becomes available to the school or the school district. Similar workflows for content publishers, especially for content review and quality check, can also be accommodated in the LCH. Other collaboration tools such as instant messaging will be enabled in future releases of the LCH.

Content and metadata storage and versioning

In the education domain, it is not uncommon to label learning objects to more than one set of metadata items. For example, a grade 9 science document that teaches about the "laws of motion" can also be labeled as a document that demonstrates concepts of force and acceleration. Both of these concepts might have different sets of metadata but they are associated with the same document. LCH manages this behavior by isolating metadata and content without compromising on the soft linkage between them. The LCH maintains a many-to-one relationship between metadata and content; i.e., one learning object can have multiple sets of metadata.

The LCH internally implements a hierarchical data model to ensure that when an instructor or a curriculum coordinator searches for a document based on its metadata, it automatically resolves any metadata links to return a list of content objects. This is especially useful for the creation of "lesson plans." A lesson plan is a logical grouping of multiple learning objects in which the goal of the lesson plan is to teach a larger concept. While creating lessons plans, teachers frequently query an LCMS system to locate appropriate content that would teach the concepts covered by learning standard instructions, and the LCH data model facilitates easy access to content.

Further, a separation of the metadata stored and the document means that any change done to either the document or the metadata should preserve the relationship between them without the intervention of the end user.

The LCH ensures that any change done to the documents or the metadata are provisioned as a version change in the system.

Content analytics for education

The Content Analytics component comprises of a series of annotators (also referred to as analyzers) that extract metadata from the educational content. This extracted metadata is stored and associated with the original document, thus allowing efficient search and retrieval of the desired content. To illustrate the usefulness of such metadata, we take the example of metadata generated by the Content Curriculum Aligner annotator. This annotator identifies what concepts, topics, and instructions are covered in a given content. The inclusion of metadata generated by the Content Curriculum Aligner in LCH, for example, simplifies the sharing and usage of content authored by teachers or sourced from open education resources. The annotator checks each document for an alignment to thousands of learning standard instructions and returns a ranked list of the best matching learning standard instructions associated with the learning content. The analyzer re-computes the metadata if the learning standard or the document content is revised.

Another set of metadata that is useful in the context of educational content is *readability* [11]. There has also been much research around understanding the readability of educational content. Different readability measures exist, such as Flesch-Kincaid readability tests [12], Gunning fog index [13], Coleman-Liau index [14], Automated Readability Index, etc. These readability tests approximate, or estimate, the years of education needed to understand a piece of writing—which helps in identifying suitable learning material for different students. The LCH also provides an annotator that automatically generates readability scores. The following subsections explain each annotator in detail.

Content curriculum aligner

In any education program, learning standards (frequently referred to as academic standards, course curriculum, etc.) define a set of skills that students should learn at different points during their learning progression. A few examples of such learning standards for K-12 learning in the U.S. educational system are as follows: Academic Knowledge and Skills (AKS), Texas Essential Knowledge and Skills (TEKS), and Common Core State Standards (CCSS). These are typically organized in a hierarchy, with instructions at the leaf nodes and a hierarchical arrangement of these instructions according to grades, subjects, courses, etc. Thus, the instructions describe the skills that students should acquire at various grades, subjects, courses, and topics. For example, "*solve linear equations*" is one such instruction, which indicates the skill a high school student should acquire in mathematics under the "linear algebra" topic in the "algebra" course. The LCH provides the ability to automatically identify the appropriate learning standard instruction for a given piece of learning content.

In order to automatically label learning content with learning standard instructions, the LCH first generates a semantic representation (also referred to as lexicons) of learning standard instructions and then uses a vector space model of the documents and the lexicons to identify the best matching instruction for each document.

To generate the semantic representation (lexicons) of instructions, LCH uses external knowledge sources such as Wikipedia**, WordNet** (a lexical database for the English language), Word vector embedding semantic representations, etc., to generate terms (expansions) that are semantically related to the key concepts mentioned in learning standards. These sets of expansions are important because, due to the short length of learning standard instructions, written for teachers, they do not include additional terms referring to concepts that work for students and implicitly described by the instruction. For example, the instruction "*relate temperature, pressure, and volume of gases to the behavior of gases*" does not refer to the concept of "*Boyle's Law*," which is likely to be important for this instruction.

The lexicons along with learning content are converted to a term frequency-inverse document frequency (*tf-idf*) based vector space representation and then used by a matching algorithm that considers the learning standard hierarchy to determine the most relevant learning standard instructions for a document. Similarity scores using the vector representation are computed for each instruction and document and scores are weighted at different levels in the hierarchy to determine the top-k best matching instructions for each document. The overall process of lexicon creation and scoring is shown in **Figure 1**. In Figure 1, DBPedia and WordNet refer to knowledge bases that can be queried to retrieve syntactic and semantic information about words; POS refers to Part of Speech tags, and "regex" denotes regular expressions that are used for pattern matching. Details of the lexicon generation and matching algorithm are described elsewhere [15].

Comprehension burden

The Comprehension Burden score [16] is a combination of four metrics that quantify how information is presented in text. These metrics are (1) Readability, (2) Concept Density, (3) Dispersion, and (4) Illustrative Richness.

Readability is used to measure the syntactic complexity of text, i.e., it measures how complicated a sentence construction is in the text. As alluded to earlier in this paper, this score is computed as an average of Flesch Reading Ease [12], Flesch Kincaid Grade Level [12],

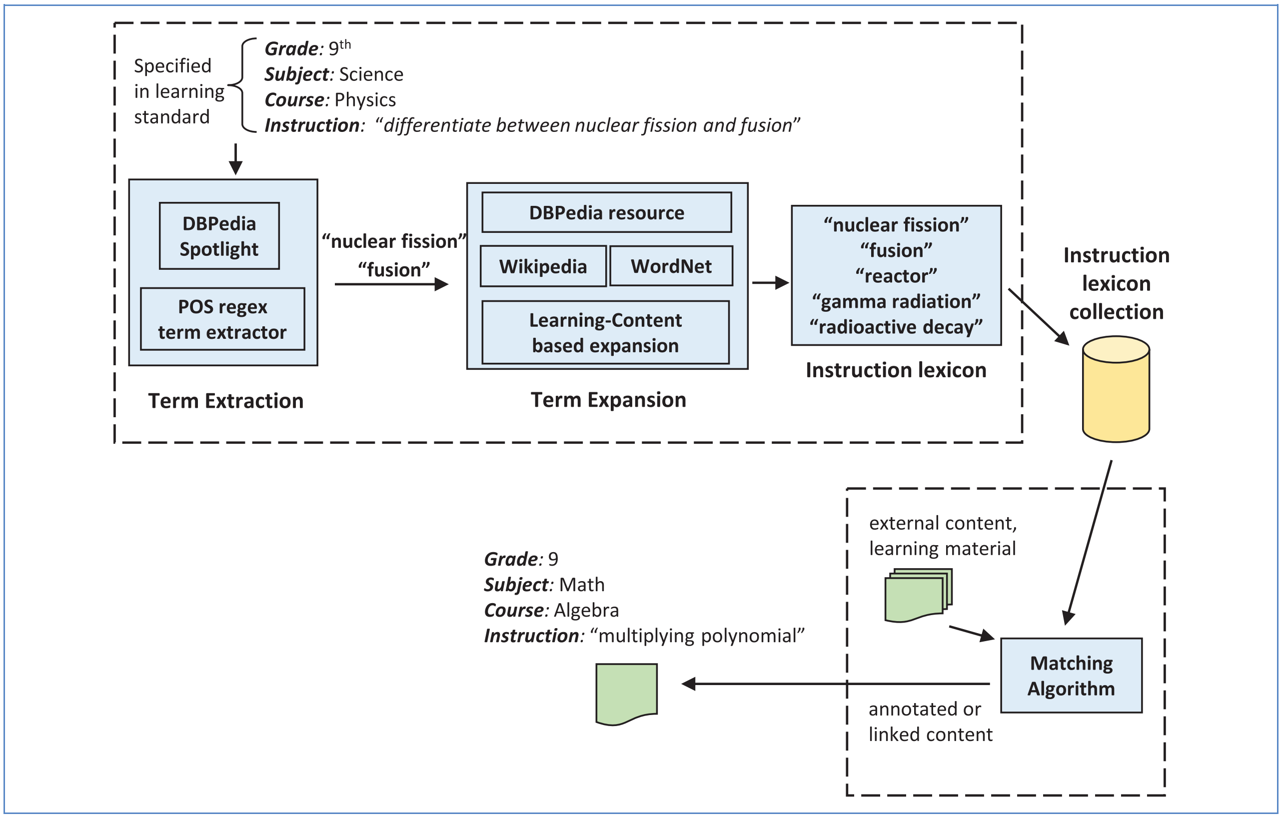

Figure 1

Content curriculum aligner architecture. Adapted from [9].

Gunning Fog Index [13], Coleman Liau Index [14], SMOG Index [17], and Automated Readability Index. The parameters used by these scores incorporate features such as sentence length, letter count, word count, average words per sentence, average syllables per word, etc. The readability score is a number between 0 and 22, with 22 indicating a very complex syntactic structure.

Concept Density is defined as the number of concepts or key terms found per unit length of text. Concepts or key terms are identified using a lexicon or external knowledge sources such as DBPedia. The higher the number of concepts in text, the less readable the text becomes for a reader to follow.

Whereas the Concept Density score measures occurrences of concepts, the Dispersion measures how far apart these concepts are, using the Wikipedia category graph. For easy-to-read, cohesive text, the dispersion of concepts in text should be gradual. Thus, dispersion is defined as the ratio of the number of unrelated concept pairs and the total number of concept pairs identified.

Illustrative Richness score quantifies the extent to which text contains illustrations, and diagrams, etc. Higher illustrative richness could result in a better understanding of content, especially in subjects such as science. These various metrics have currently been studied and applied on English language content.

Search

The LCH provides a search component that enables users to retrieve relevant learning contents from the LCH repository. The system provides a search mechanism where one can specify a keyword query. Learning content containing the specified keywords is returned as a search result. In addition, learning contents can be searched using a set of metadata like "Grade," "Course," "Topic," etc. The search component can also be used by any LMS to retrieve learning contents through search APIs exposed by the LCH system.

In the future, the LCH will support other forms of complex search mechanisms that will allow users to quickly identify relevant documents. One direction that is being currently investigated is the providing of a mechanism of search-result clustering, i.e., the result set returned by keyword and/or attribute search is organized around fixed dimensions (facets) or dynamically

Table 1 Topic-level labeling performance for dataset A. (MMA: Minimal match accuracy; FMA: Full match accuracy; MRR: Mean reciprocal rank.) Adapted from [9].

Configuration	*High school mathematics*							*High school science*						
	MMA (%)		*FMA* (%)		*Recall*		*MRR*	*MMA* (%)		*FMA* (%)		*Recall*		*MRR*
	@5	@10	@5	@10	@5	@10		@5	@10	@5	@10	@5	@10	
NoExpansion	75.0	83.3	66.6	66.6	0.68	0.73	**0.82**	79.8	88.4	76.9	86.3	0.77	0.86	0.72
WordNet	58.3	83.3	41.6	75.0	0.52	**0.84**	0.72	76.9	84.1	73.3	81.2	0.74	0.82	0.68
Wikipedia	**83.3**	**83.3**	**75.0**	**75.0**	**0.79**	0.79	0.70	**87.7**	**90.6**	**84.8**	**88.4**	**0.85**	**0.88**	**0.74**
WordEmbedding	83.3	83.3	66.6	66.6	0.79	0.79	0.74	79.1	85.6	75.5	82.7	0.76	0.83	0.69
Wiki+WordNet	58.3	75.0	41.6	66.6	0.47	0.73	0.66	80.5	87.7	77.6	84.8	0.78	0.85	0.73
Wiki+WordEmbedding	75.0	83.3	66.6	75.0	0.73	0.84	0.68	81.2	85.6	77.6	82.0	0.78	0.83	0.72

Table 2 Topic level and instruction level (AKS) labeling performance for dataset B. (MMA: Minimal match accuracy; FMA: Full match accuracy; MRR: Mean reciprocal rank.)

Grade-Subject	*#Docs*	*AKS*							*Topic*						
		MMA (%)		*FMA* (%)		*Recall*		*MRR*	*MMA* (%)		*FMA* (%)		*Recall*		*MRR*
		@5	@10	@5	@10	@5	@10		@5	@10	@5	@10	@5	@10	
HighSchool-Math	14	64.2	85.7	28.5	50	0.54	0.71	0.56	92.8	100	85.7	92.8	0.87	0.93	0.61
HighSchool-Science	89	79.7	84.2	2.2	5.6	0.27	0.29	0.63	93.2	96.6	80.9	86.5	0.81	0.88	0.81
8^{th}-Math	79	60.7	60.7	12.6	12.6	0.24	0.24	0.35	97.4	100	74.6	100	0.79	1	0.49
8^{th}-Science	40	87.5	90.0	7.5	12.5	0.31	0.33	0.61	95.0	100	72.5	100	0.81	1	0.73
7^{th}-Math	23	95.6	95.6	34.7	43.4	0.55	0.57	0.73	100	100	73.9	100	0.8	1	0.93
7^{th}-Science	40	97.5	97.5	12.5	12.5	0.54	0.55	0.88	97.5	100	98.0	100	0.98	1	0.95
6^{th}-Math	38	68.4	81.5	26.3	34.2	0.31	0.38	0.32	92.1	100	71.0	100	0.79	1	0.58
6^{th}-Science	49	91.8	93.8	28.5	42.8	0.59	0.66	0.79	100	100	97.9	100	0.98	1	0.97
5^{th}-Math	23	100	100	26.0	30.4	0.58	0.61	0.95	100	100	100	100	1	1	1
4^{th}-Math	51	88.2	98.0	47.0	52.9	0.66	0.69	0.90	100	100	100	100	1	1	0.98

grouped into latent dimensions that are discovered based on the characteristic of the search result.

Evaluation

In this section, we describe experiments demonstrating the performance of one of our annotators, namely the Content Curriculum Aligner annotator. We also discuss our experience in deploying the LCH system with a prominent school district in United States.

To evaluate the performance of the Content Curriculum Aligner annotator, two data sets were used. These are referred to as Dataset A and Dataset B. Dataset A consisted of 179 educational documents for two subjects, namely high school math and science. Dataset B consisted of a total of 446 math and science educational documents from different grades. For all documents in Dataset A, we had ground truth (correct labels) until the topic level in the AKS hierarchy, while for Dataset B the ground truth was until the AKS instruction level was available.

Table 1 and **Table 2** show our system's performance on these two datasets. The Minimal Match Accuracy (MMA) is a precision-oriented measure that is based on the number of documents with at least one system assigned label being present in the ground truth. The stricter metric called the Full Match Accuracy (FMA) requires that all instruction labels assigned by the algorithm be present in the ground truth. We also evaluate the system using the Mean Reciprocal Rank (MRR), which is the average of the reciprocal rank of the highest ranked correctly assigned system label for each document.

The experiments were performed using different methods for generating instruction lexicons. These lexicons were then used by the content matching method as described earlier to generate the top scoring labels for each document. Five different sources were used to generate the lexicons, namely, WordNet, Wikipedia, Word vector embedding, and combinations of Wikipedia with WordNet and Word embedding. As can be seen from Table 1, where the bold font indicates the highest value in each metric, our system has a topic level full match accuracy of 84.89% for science and 75.00% for mathematics, while the full match accuracy @10 (precision in the top 10 labels assignment by the system) is 88.48% for science and 75.00% for mathematics. Further details on the evaluation measures, lexicon generation algorithm, and experiments can be found elsewhere [9].

From Table 2, it can be seen that our system has comparable results on Dataset B. The Instruction level accuracies can be seen in the column titled "AKS." Both Minimal match and full match accuracies at the instruction level have been shown. The results from the evaluation demonstrate that our analyzers are generic and report consistent performance on different data sets.

We next share our experience in deploying the LCH system at GCPS (Gwinnett County Public School), a school district in the State of Georgia. The primary users of the pilot were 11 math and science teachers from middle school and high school. All of these teachers were fairly experienced in the use of IT systems in classrooms, including the use of LMSs. The content uploaded by the GCPS teachers onto LCH was retrieved and used within interventions assigned to students using an application called Personalized Learning Pathways (PLP) [18], which in turn uses student risk prediction system [19, 20] to identify students in need of personalized interventions. PLP uses LCH as the content repository at the back-end, and lets users search for content by selecting the AKS curriculum standard instructions for which an intervention needs to be created. These instructions are typically those for which a student has demonstrated certain deficiencies. The AKS tags generated for each learning resource by LCH is used to retrieve a list of potential intervention content. The teacher who is creating an intervention then manually reviews the search results in PLP and makes appropriate selections for the intervention. Thus, the metadata assigned by LCH is a core capability used to design effective personalized interventions for students. Overall, 259 interventions have been created in this way during a PLP pilot study at GCPS. For details on PLP and the pilot, the interested reader may refer to [18].

At the end of the pilot, a teacher survey was conducted, which provided several useful recommendations to enhance LCH's metadata capabilities. Specifically, the following features were requested:

- Ability to preview content in LCH, along with title and other descriptors such as content source, submitter's name, submission date, etc.
- Type of learning resource (e.g., video, read and respond, multiple choice, short answers, etc.).
- Expected completion time.
- Student engagement score.
- Teacher-assigned effectiveness score (and age appropriateness).
- System-assigned effectiveness score (e.g., based on usage and outcomes).
- Usage data, e.g., over a period (monthly, semester-wise, or yearly), by grade, repeated usage by same user (student or teacher), etc.
- Content usage policy, addressing any known restrictions or constraints in usage.

These recommendations demonstrate the wide variety of educational content metadata that a small group of pilot users expect to find useful. This is certain to grow as LCH is used by other groups or in other contexts. They validate the need for a smart metadata generation and management system for learning content like LCH. Some of these recommendations are helping shape the LCH research agenda for the future.

Conclusion and future work

In this paper, we described a platform called LCH for managing and analyzing educational content. Beyond content management functionalities—such as access and version control, document workflow support, etc.—this platform provides support for advanced search and content analytics capabilities. We provided details of some of the content analyzers that are currently part of the LCH, with specific details on the Curriculum Linking analyzer. The LCH is designed in such a way that it allows rapid deployment of new content analyzers and support for ingestion of content from varied data sources, as well as industrial-strength user and access management. We shared details on each of these aspects in the paper. Our experience in deploying this system with a U.S. School district was also shared.

In the future, we plan to support (1) storage, analysis, and retrieval of multimedia objects, (2) faceted search, (3) richer content analysis (e.g., parsing mathematical equations and chemical formulas), (4) extended API support and toolkits for development of educational applications on LCH, and (5) expanded breadth of subjects within learning standards we handle, as well as the breadth of languages in which learning content can be authored and cataloged.

Acknowledgments

Ron Campbell, John P. Costantino, James V. Stembridge, John Hunt, Doug Billings, Loai Said, Christopher Matthew, and Kevin Ho of IBM Global Business Services Analytics Practice, along with IBM Software Group and IBM Interactive, did extensive work with Gwinnett County Public Schools (USA), allowing a thorough test of

LCH's handling of all the middle-school and high-school math and science content used in this very large and ambitious school district. Lisa Burleson shared an experienced teacher's perspective, provided input on metadata quality and use cases, and evaluated our comprehension burden analyzer. Bob Schloss of the IBM Thomas J. Watson Research Center worked with LMS vendors and helped us implement IMS standards support.

*Trademark, service mark, or registered trademark of International Business Machines Corporation in the United States, other countries, or both.

**Trademark, service mark, or registered trademark of Wikimedia Foundation or Princeton University in the United States, other countries, or both.

References

1. W. R. Watson and S. L. Watson, "An argument for clarity: What are learning management systems, what are they not, and what should they become?" *TechTrends*, vol. 51, no. 2, pp. 28–34, Mar. 2007.
2. *Learning Standards*. [Online]. Available: http://edglossary.org/learning-standards/
3. "Academic knowledge and skills learning standard," Gwinnett County Public Schools, Suwanee, GA, USA, 2014. [Online]. Available: https://publish.gwinnett.k12.ga.us/gcps/home/public/parents/content/general-info/aks
4. X. Ochoa, K. Cardinaels, M. Meire, and E. Duval, "Frameworks for the automatic indexation of learning management systems content into learning object repositories," in *Proc. ED-MEDIA*, Montreal, QC, Canada, 2005, pp. 1407–1414.
5. *Competency-Based Learning or Personalized Learning*. [Online]. Available: http://www.ed.gov/oii-news/competency-based-learning-or-personalized-learning
6. "Competency-Based Education (CBE)," Educause Library, Louisville, CO, USA, [Online]. Available: http://www.educause.edu/library/competency-based-education-cbe
7. M. R. Weise, "Got skills? Why online competency-based education is the disruptive innovation for higher education," *Why IT Matters to Higher Education, Educause Review*, 2014. [Online]. Available: http://www.educause.edu/ero/article/got-skills-why-online-competency-based-education-disruptive-innovation-higher-education
8. "Personal Learning Plan," Glossary Educ. Reform, Portland, ME, USA. [Online]. Available: http://edglossary.org/personal-learning-plan/
9. "IMS content packaging specification primer," IMS Global Learning Consortium, Apple Valley, MN, USA, Version 1.2, 2007. [Online]. Available: http://www.imsglobal.org/content/packaging/
10. D. Ferruci and A. Lally, "UIMA: An architectural approach to unstructured information processing in the corporate research environment," *Nat. Language Eng.*, vol. 10, no. 3/4, pp. 327–348, Sep. 2004.
11. E. Dale and J. S. Chall, "The concept of readability," *Elementary English*, vol. 26, no. 23, pp. 24–29, Jun. 1949.
12. J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for navy enlisted personnel," Chief Naval Tech. Training: Naval Air Station Memphis, Millington, TN, USA, Res. Branch Rep. 8-75, 1975.
13. R. Gunning, "*The Technique of Clear Writing*," New York, NY, USA: McGraw-Hill, 1952, pp. 36–37.
14. M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring," *J. Appl. Psychol.*, vol. 60, no. 2, pp. 283–284, Apr. 1975.
15. D. Contractor, K. Popat, S. Ikbal, S. Negi, B. Sengupta, and M. K. Mohania, "Labeling educational content with academic learning standards," in *Proc SIAM Int. Data Mining*, Vancouver, BC, Canada, 2015, pp. 1–2.
16. R. Agrawal, S. Chakraborty, S. Golapudi, A. Kannan, and K. Kenthapadi,"Empowering authors to diagnose comprehension burden in textbooks," in *Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, 2012, pp. 967–975.
17. McLaughlin and G. Harry, "7 Grading—A new readability formula," *J. Reading*, vol. 12, no. 8, pp. 639–646, 1969.
18. V. K. Reddy, L. Said, B. Sengupta, M. Chetlur, J. P. Costantino, A. Gopinath, S. Flynt, P. Balunaini, and S. Vedula, "Personalized Learning Pathways: Enabling intervention creation and tracking," *IBM J. Res. Dev.*, vol. 59, no. 6, Paper 4, pp. 4:1–4:14, 2015.
19. S. Ikbal, A. Tamhane, B. Sengupta, M. Chetlur, S. Ghosh, and J. Appleton, "On early prediction of risks in academic performance for students," *IBM J. Res. Dev.*, vol. 59, no. 6, Paper 5, pp. 5:1–5:14, 2015.
20. A. Tamhane, S. Ikbal, B. Sengupta, M. Duggirala and J. Appleton, "Predicting student risks through longitudinal analysis," in *Proc. 20th ACM SIGKDD Conf. KDD*, New York, NY, USA, 2014, pp. 1544–1552.

Received February 4, 2015; accepted for publication March 1, 2015

Danish Contractor *IBM Research, New Delhi, India (dcontrac@in.ibm.com).* Mr. Contractor is a member of the Education Transformation Research Group at IBM Research - India. He has more than 7 years of research and development experience at IBM. During this time, he has been part of product development teams such as IBM eDiscovery Analyzer and IBM Case Manager, as well a member of Research teams working on social media analytics, education research, and unstructured information management. His research interests include machine learning and natural language processing, and he has published 14 papers in top-tier conferences and has applied for 8 patents. Mr. Contractor holds an M.Phil. degree in advanced computer science from the University of Cambridge and a Bachelor's degree in computer engineering from the Netaji Subhas Institute of Technology, New Delhi.

Sumit Negi *IBM Research, Bangalore, India (sumitneg@in.ibm.com).* Mr. Negi is a Senior Software Engineer in the Education Transformation Group at IBM Research - India. He has more than 12 years of research and development experience. Over these years, Mr. Negi has been responsible for the conceptualization, design, and development of information management and analytics solutions for clients from different industry verticals. His research interests are in the area of machine learning, natural language processing, and social networks. He has authored more than 30 papers in top-tier ACM and IEEE conferences. Mr. Negi received his M.B.A. degree from Faculty of Management Studies (New Delhi), his M.S. degree in computer science from IIT (Delhi), and a Bachelor's degree in electronics and communication from NIT (Durgapur, West Bengal).

Kashyap Popat *IBM Research, Bangalore, India (kaspopat@in.ibm.com)*. Mr. Popat is an IBM Blue Scholar in the Smarter Education group at IBM Research. He received an M.Tech. degree in July, 2013 from Indian Institute of Technology, Bombay (IITB). His research interests include natural language processing, text analytics, machine learning, and data mining.

Shajith Ikbal *IBM Research, Bangalore, India (shajmoha@in.ibm.com)*. Dr. Ikbal is a Research Scientist at IBM Research - India. He was part of Smarter Education Group while performing this work. His research interests include speech recognition or understanding, machine learning, data mining, and text analytics.

Balunaini Prasad *IBM India Software Lab, Bangalore, India (balunaini@in.ibm.com)*. Mr. Prasad is a developer in IBM Software Group at India Software Labs. He develops IBM Case Manager Builder. Prior to this product, he had worked in teams associated with IBM eDiscovery Analyzer, IBM Content Analyzer, DB2*, and AIX* development. He holds a Master's degree in electrical engineering from Indian Institute of Technology, Kanpur.

Sandeep Vedula *IBM India Software Lab, Bangalore, India (vvedula@in.ibm.com).* Mr. Vedula works as a Software Engineer in the product group of IBM India Labs. He has 6 years of experience in software development in the areas of analytics, content management, online learning, and supply chain management. Mr. Vedula has a Master's degree in artificial intelligence from Central University, Hyderabad, and keeps keen interest in big data and data science technologies.

Sreekanth Kakaraparthy *IBM India Software Lab, Bangalore, India (sreekanthkl@in.ibm.com)*. Mr. Kakaraparthy is a Senior Technical Leader in the IBM Analytics division at the IBM India Software Lab. In 1997, he received a Bachelor's degree in engineering from Andhra University. He subsequently joined IBM, where he has worked on developing products and industry solutions in enterprise content management (ECM) areas. He worked extensively with clients and partners and developed innovative solutions addressing the requirements from the industry. He received an IBM Outstanding Technical Achievement Award for client value, and has many other awards and citations. He is author or coauthor of two patents and two technical papers.

Bikram Sengupta *IBM Research - India, India, Bangalore (bsengupt@in.ibm.com)*. Dr. Sengupta is a Senior Technical Staff Member at IBM Research - India and leads the Smarter Education Group, where he oversees work in the area of tools and techniques for personalized education. Prior to this, he led the Services Analytics Group at IBM Research - India from 2008 to 2012. In the past, he has worked extensively in the areas of global software development, formal techniques for system analysis and design, and systems management. Dr. Sengupta received a B.E. degree in computer science and engineering from Jadavpur University, India, in 1998 and M.S. and Ph.D. degrees in computer science from Stony Brook University, New York, USA, in 2000 and 2003, respectively.

Vijay Kumar *IBM India Software Lab, Bangalore, India (vjkumar@in.ibm.com).* Mr. Kumar is a Senior Manager leading the ECM (enterprise content management) solutions team at the IBM India Software Lab—for IBM Case Manager for Investigations, IBM Counter Fraud Management, and IBM Watson Foundation for Education.