InVideo: An Automatic Video Index and Search Engine for Large Video Collections

Shuangbao Paul Wang

Metonymy Corporation
Fairfax, Virginia 22030
Email: paul.wang@computer.org

Xiaolong Cheng

Department of Computer Science
George Washington University
Washington, DC 20052
Email: xiaolongcheng@gwu.edu

Carolyn Maher

Department of Psychology
Rutgers University
New Brunswick, New Jersey 08901
Email: carolyn.maher@csu.rutgers.edu

William Kelly

Metonymy Corporation
Fairfax, Virginia
Email: william.kelly@metonymylabs.com

***Abstract*—In this paper, we present a novel system, inVideo, for automatically indexing and searching videos based on the keywords spoken in the videos and the content of the video frames. Using the highly efficient video indexing engine we developed, InVideo is able to analyze videos using machine learning and pattern recognition without the need for initial viewing by a human. The time-stamped commenting and tagging features refine the accuracy of search results. The cloud-based implementation makes it possible to conduct elastic search, augmented search, and data analytics. Our research shows that inVideo presents an efficient tool in processing and analyzing videos and increasing interactions in video-based online learning environment. Data from a cybersecurity program with more than 500 students show that applying inVideo to current video material, interactions between student-student and student-faculty increased significantly across 24 sections program-wide.**

***Index Terms*—video processing; video index; big data; learning analytics.**

I. Introduction

Big data analytics are used to collect, curate, search, analyze, and visualize large data sets that are generated from sources such as texts (including blogs and chats), images, videos, logs, and sensors [1]. Video data is a major format of unstructured data, and should be an indispensable area of big data analytics. However, most analytics tools are only effective in analyzing structured data. Due to the nature of the special file format, traditional search engines hardly penetrate into videos, and therefore video indexing becomes a problem [2]–[13].

Videos contain both audio and visual components, and neither of these components is text based. To understand a video, viewers must actually play it and use their eyes and ears to analyze the sounds and visuals being presented to them. Without watching a video, it is hard to glean information from its content or even know whether there is information to be found within. Existing search engines and data analytics tools such as Google, SAS, SPSS, and Hadoop are effective only in analyzing text and image data. Video data, however, are difficult to index and therefore difficult to analyze.

In education, video presents a large opportunity for both classroom and online education [14]. In addition, video is a great teaching format because it can both be more enjoyable and more memorable than other instruction formats [15]. Furthermore, video instruction allows for students to work at their own pace, for teachers to be able to teach more students, and for more reusable teaching materials to be available when compared to an in-person lecture. MOOC creators realize the many benefits of video, as evidenced by the prevalence of video in MOOCs. Many MOOCs focus on video files for the bulk of their instructional material so it is clear that the MOOCs of the future must also focus on videos.

InVideo [16], developed under a US Department of Education grant, is able to analyze video content (language and video frames) prior to initial close researcher review of the video. A highly efficient video indexing engine can analyze both language and video frames based on natural language and referent objects. Once a video is indexed, its content becomes searchable and statistical analysis as well as qualitative analysis are possible. Commenting and tagging add a layer of hyper-information and therefore increase the accuracy of the transcript, which was automatically extracted from the video by the inVideo tool. The indexing technology is especially useful in mining video data in large video collections. inVideo also has an automatic caption system that can transcribe the words spoken in the video. Instructors can use the tool to construct in-place video quizzes for assessments.

Learning is an integration of interaction. The interaction might exist between learners and instructors or between learners and computers. While the traditional approach would be to analyze grades at the end of the semester, this lacks the benefits that come from interactions that occur during the course [17]. As an increasingly large number of educational resources

 ISBN: 978-1-61208-559-3

move online, analyzing interactions between students and online course material is becoming more important. Many learning management systems (LMS) have built-in learning analytics tools to look into the data [18]–[20]. Due to the limitation of the data gathering and indexing, the built-in tools are generally not sufficient in assessing study outcomes, especially for video content.

II. Related Work

Automatic video index and search have widely applications in education, public security, and many other video-intensive areas.

An airport traffic and security monitoring system constantly index videos gathers from surveillance cameras and search the suspect based on the graphical and textual information provided by the authority [9].

A video content indexing and retrieval tool index digital videos automatically on a 34 hours of TV news broadcast. The sampled frames are then used in providing the basis for various analysis [21].

Big data and learning analytics can become part of the solutions integrated into administrative and instructional functions of higher education [22]. Traditional face-to-face instruction supports traditional data-driven decision-making process. Videos as a form of big data are more extensive and especially time-sensitive learning analytics applications. It is important that instructional transactions are collected as they occur.

Learning analytics can provide powerful tools for teachers in order to support them in the iterative process of improving the effectiveness of their course and to collaterally enhance their students performance [23]. Dyckhoff developed a toolkit to enable teachers to explore and correlate learning object usage, user behavior, as well as assessment results based on graphical indicators. This learning analytics system is able to analyze data such as time spent, areas of interest, usage of resources, participation rates and correlation with grades data and visualize them using a dashboard. However, the system is unable analyze the interactions between students and the online learning systems on videos.

In order to analyze videos for various applications, we have developed a video index engine to look at every word spoken in the video and categorize it using our custom index algorithm. In addition, a content-based pattern recognition engine can search individual frame of the video to recognize objects and individuals being displayed. The collaborative commenting, tagging, and in-place quizzes make videos more accessible and also increase the accuracy of the search engine [7], [8], [10].

III. Video Indexing and Search Algorithm

Videos are a different data type than text and images, in that they are unstructured data. Traditional search engines are mostly text based, with a few tools that allow for searching of images. In order to index a video, a search engine needs

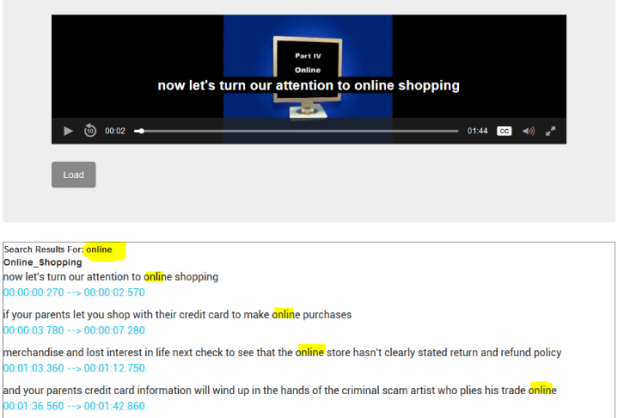

Fig. 1. Analyzing Videos by Keywords

to extract meaningful language from the audio and convert it to text, while simultaneously converting the visual frames into a series of images that can be used to recognize persons and objects in the video. This is an extremely difficult task, given that videos are a compound format. Not only are the audio and visual components integrated, but also within each of these components there is a blend of information being presented in a manner that cannot be distinguished as easily by a computer as by a human brain. For example, the audio of the file may contain speech, music, and background noises that a computer will have a hard time recognizing and analyzing.

A. Automatic Indexing Algorithm

The video indexing engine uses the vector space model to represent the document by a set of possible weighted content terms. The weight of the term reflects its importance in relation to the meaning of the document [24].

After calculating the normalized frequency of a term in the document, the weight to measure the relative importance of each concept or single term is obtained. The automatic index algorithm then calculates the final position in n-dimensional space. The result is to be used for generating search results or visualization.

B. Searching Videos by Keywords

Video search involves two steps: analyzing by keywords and analyzing by image references. When a keyword is entered, the system looks through the indexed audio transcript to see if there is a match. An image reference may refer to either a picture or keywords that describe an object in the video using an appropriate semantic space. Video clips whose language contains the keywords will be retrieved. Figure 1 shows how indexed videos can be searched using keywords in the spoken language.

C. Searching Videos by References

Searching videos by references examines the frames of the video to see if the given picture or keyword is found. If the

 ISBN: 978-1-61208-559-3

Fig. 2. Analyzing Videos by Image References using the CBIR Algorithm

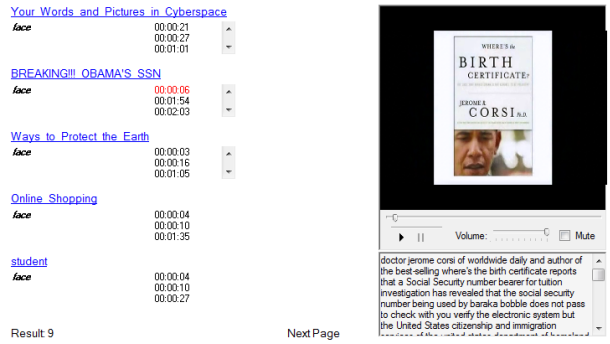

Fig. 3. Analyzing videos by Keyword References using a Knowledge Tree

reference is a picture, then the system uses a Content Based Image Retrieval (CBIR) algorithm to find the match frames and return the video clips that contain the reference picture. Figure 2 shows the image-based CBIR algorithm that retrieves the video frames corresponding to the reference picture (at the bottom).

If the reference is a keyword (e.g. credit card) then the system uses a knowledge tree to find matches in the video. If one video frame contains an object matching the features associated with the keyword, the section of the video is returned. Figure 3 shows how a search for the keyword credit card will retrieve the video frames that contain objects as credit cards.

D. Searching Videos with Multiple Languages

Sometimes multiple languages may be found in videos. Transcribe engines normally only work in one language or in closely-related languages. For other languages, a different transcribe engine may be required. InVideo addresses this problem by allowing videos with different languages to be searched from a single user interface. The inVideo system does not translate between languages. It only transcribe based on the language of original videos. For example, a Chinese video will result in a transcript in Chinese. Figure 4 shows the indexing engine properly analyzing the Chinese language.

Fig. 4. Analyzing Videos with Different Languages

When entering the word student in Chinese, the video search engine will locate that term in the transcript and return the corresponding frames. Currently, there are multiple languages that can be analyzed by the inVideo system, with more to be added.

E. Elastic Search and Content-aware Elastic Search

Elastic search makes the inVideo application capable of searching video data across distributed environment with HTTP protocol and schema-free JSON documents. Elastic search makes it possible to expand the community by deploying a pluggable cloud architecture, configurable automatic discovery of cluster nodes, persistent connections, and load balancing across all available nodes. Video collections under ACE are no longer restricted to a particular video collection. More importantly, there is no need to move other video collections in a centralized site, which is merely feasible anyway.

Indexed videos can be searched by keywords. When a keyword is entered, the cloud system searches its generated transcript of the audio files to find matches. Video clips whose audio track contains the keywords are retrieved.

The enhanced elastic search partitions videos into individual frames. Thus, users also have the ability to search a video by examining the video frames to see if a given object is found. The enhanced elastic search algorithm matches contents in the frames and return the video clips that contain the reference objects. This process allows us to combine images and words to create hybrid metadata.

For instance, in mathematics education, there is interest in studying the representations students make. The CBIR algorithm allows us to search the videos for particular representations a student would make, for example, images of rods, blocks, tree diagrams. Students may construct these models without talking about them, and thus an audio transcript would miss it. By being able to search the frames of the video, a user interested in tracing a students construction or explanation of a representation can query the database. If a frame of the video contains an object matching the particular representation, this section of the video is returned.

 ISBN: 978-1-61208-559-3

F. Machine Learning and Cloud-based Data Analytics

It should be noted that the goal of the software is not to reproduce the word-for-word accuracy of a human-generated transcript of video files. Rather, the goal is to determine the extent to which voice-to-text analysis and image analysis are able to retrieve desired sections of the video for enriched human analysis. The critical task is for the system to identify a sufficient amount of relevant hits that pertain to the search reference terms or images. Search queries have advanced from keywords to natural language. The inVideo uses artificial intelligence and machine learning technologies to analyze videos and use big data analytics tool to explore, index and visualize videos. InVideo also has security features that secure data and communication in the cloud and protect privacy [2]–[6], [11], [12].

IV. Collaborative and Interactive Learning

Automatically generated video transcripts may have accuracy problems. Besides, the vast numbers of videos in MOOCs make them impossible to be retrieved correctly with just one or a few simple keywords. To solve these problems, we have implemented a collaborative filtering mechanism including commenting, tagging, and in-place quizzes. These features improve accuracy and increase interactions between students and the online learning systems. With collaborative filtering, learning resources retrieval on MOOC systems is greatly improved, and better student achievement is therefore expected.

A. Collaborative Filtering

Collaborative filtering is a process of improving accuracy of the automatic indexing algorithm by leveraging user feedback. This is popular on websites that have millions of users and user-generated content. Users are able to create time-stamped comments on videos. These comments can be hidden or made public so that someone else who views the video can see the comment at a specific time. These comments help increase accuracy of the search tool and transcript and enhance interactions in online learning.

Tagging on videos is another implementation in the inVideo system. It is to attach keyword descriptions to identify video frames as categories or topic. Videos with identical tags can then be linked together allowing students to search for similar or related content. Tags can be created using words, acronyms or numbers. It is also called social bookmarking.

A search term usually yields many related results, which in many cases are hard to differentiate. Commenting and tagging add additional information, refine the knowledge and increase the video search accuracy. At the time when information is exponentially growing, these features are extremely helpful for students to obtain the knowledge with the least amount of time.

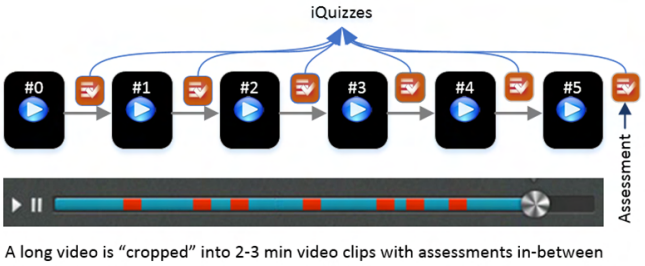

Fig. 5. Transform Linear Videos into Interactive Learning Objects

B. In-place Assessment with iQuiz

Internet computing has the advantage of employing powerful CPUs on remote servers to provide applications across the network. inVideo comes with an Internet Computing-based video quiz system (iQuiz) to utilize the computational power of remote servers to provide video quiz services to users across the Internet. Currently, videos are mostly non-interactive, therefore, there are no interactions between students and the learning content. Students view videos either online or download them to their personal devices. There is no way for educators to know whether a student has understood the content or even to know whether the student has viewed the video or not.

iQuiz can be used to assess learning outcomes associated with video study. Quizzes can be embedded into videos at any place where an instructor wants to assess the outcome of the students study. iQuiz runs as a service on servers. This enables users to execute this resource-intensive application with personal computers or iPads, which would not be possible otherwise.

Instructors can enter into the authoring mode where they can write quizzes by indicating the start and stop positions on the video and adding questions. Video quizzes are stored in XML format, and are automatically loaded while students are watching the video in the learning mode. Answers to the quizzes, either correct or incorrect, are also stored in the XML database for immediate assessments. Assessment of adaptive learning on videos provides better outcomes for students than the traditional video content study with little or no feedback [25].

C. Transform Linear Videos into Interactive Learning Objects

Video are linear in nature. It is hardly interactive nor does it contain branches. Using the inVideo tool, classical videos can be transformed into a series of video clips with assessments in between and at the end. So the video-based learning material becomes interactive. Figure 5 shows a test we conducted that turned a 46-minute video into six selected 2-3 minute video clips. The red segments on the stage bar are the samples. So it is clear that not all videos content was used in the samples.

 ISBN: 978-1-61208-559-3

V. Experimental Results

To test the inVideo system, we selected the 20 most recent videos from National Science Digital Library (NSDL) in cybersecurity and used the inVideo tool to extract keywords that appeared in the transcripts. From this set, we selected the top two ranked keywords: Target (data breach) and encryption (using encryption to secure data). We were confident that those two keywords made good discussion topics that could increase classroom interactions.

As a result, we added two discussion topics to the spring 2014 Masters of Science in Cybersecurity program (24 class sections with each section has 25 student on average).

Videos lack interactions between learners and the online learning environment. Even worse, videos above a certain length will likely never be watched at all because students cannot easily determine what content is within it or how to locate that content. To address this issue, we used the inVideo tool to index the content and break the large videos into a series of small video clips. By doing so, we made it possible for students to watch short video clips covering individual key concepts directly, while retaining the ability to view the whole video if necessary. This served to not only increase student interest and engagement in the lesson, but also more importantly, to improve their ability to comprehend and retain information.

Student responses and interactions can be used as a proxy for their degree of engagement with any particular part of the course. As one example of how the inVideo indexing served to increase this measure, consider Week 2 of the class. In our assessment of past offerings (pre-inVideo) we discovered that this part of the course is a quiet week, because the individual assignment starting in the week will not be due until Week 8. This meant that the interactions in the classrooms dropped significantly from Week 1. Based on this assessment, we decided to use the inVideo intervention in an attempt to generate more interactions during Week 2 of the course.

Our initial observation of one class was very promising; the total number of responses, defined as each posting after viewing a video clip, for Week 2 reached sixty-eight, as compared to only two for the same week in the previous semester. This initial finding encouraged us to investigate the results for all twenty-four sections program-wide. Figure 6 shows the number of responses for the 24 sections comparing Fall 2013 to Spring 2014 during Week 2.

For the research we conducted, Week 2 student responses across the 24 sections were almost seven times higher during Spring 2014 (1,129 responses) than during Spring 2014 (164 responses).

For the cybersecurity online/hybrid class, we have five graded discussions, one individual assignment, one team assignment, and two lab assignments. Two more hands-on exercises (labs) have been added since Spring 2014. Data from the team projects, using the same intervention method, show that student-student and student-faculty interactions were 6.5

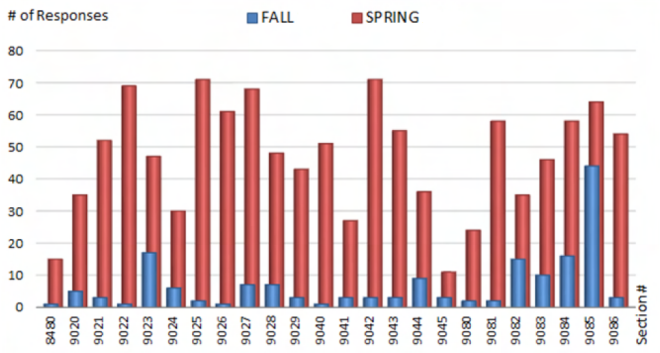

Fig. 6. Number of Responses for 24 Sections - Week 2 Discussions

times greater for the courses with the inVideo intervention (104 responses compared to 16 responses). We also measured student performance against desired learning outcomes. The average grades on both team projects and final grades was higher in Spring 2014 than in Fall 2013. Here we see that the index and data analytics tool inVideo, in combination with just-in-time assessment and intervention, improved learning outcomes.

Based on our finding, we are in the process of breaking up every large learning module into several learning objects using inVideo. The new competency-based learning objects will be used to construct the knowledge cloud. These new learning modules will consist of many competency-based learning objects, and will be more interactive, rational, and accessible.

We will use inVideo to expand the scope of this research to other activities in courses within the cybersecurity program. This tool could also be useful to courses in other disciplines. Using the inVideo tool, linear videos are transformed into a series of interactive learning objects. This is vital in an online learning environment where interactions and learning outcomes are valued the most.

VI. Conclusion and Future Work

This paper discussed a novel video index and analytics tool to analyze video data. Video indexing engines analyze both audio and visual components of a video, and the results of this analysis provide novel opportunities for search. To improve accuracy, we can either improve the transcribe engine, analyze video frames better where there is no audio, or crowd-source accuracy through collaborative filtering. For transcription accuracy, one potential accuracy improvement can come from using a self-learning artificial intelligence (AI) system that could be taught to recognize certain accents or languages. The process or requirements for instituting such a system and the magnitude of the improvement in accuracy are to be studied in the future.

At present, inVideo tool is only limited to analyze native (non-streaming) videos. We will continue our research on analyzing live videos and streaming videos and make inVideo available to broader video collections and applications.

 ISBN: 978-1-61208-559-3

ACKNOWLEDGEMENT

This research is funded in part by grants from US National Science Foundation (NSF) [EAGER-1419055 and DGE-1439570].

REFERENCES

[1] K. Bakshi, "Considerations for big data: Architecture and approach," 2012, pp. 1–7.

[2] S. Wang, "Dual-data defense in depth improves scada security," *Signal*, pp. 42–44, 2016.

[3] S. Wang and W. Kelly, "Smart cities architecture and security in cybersecurity education," *The Colloquium of Information Systems Security Education (CISSE)*, 2017.

[4] P. Wang and W. Kelly, "A novel threat analysis and risk mitigation approach to prevent cyber intrusions," *Colloquium for Information System Security Education (CISSE)*, vol. 3, pp. 157–174, 2015.

[5] S. Wang, A. Ali, and W. Kelly, "Data security and threat modeling for smart city infrastructure," 2015, pp. 1–6.

[6] S. P. Wang and R. S. Ledley, *Computer Architecture and Security*. Wiley, 2013.

[7] S. Wang, W. Kelly, and J. Zhang, "Using novel video indexing and data analytics tool to enhance interactions in e-learning," 2015, pp. 1919–1927.

[8] S. Wang, A. Ali, J. Zhang, and W. Kelly, "invideo - a novel big data analytics tool for video data analytics and its use in enhancing interactions in cybersecurity online education," vol. 60, 2014, pp. 321–328.

[9] S. Wang and J. Zhang, "A video data search engine for cyber-physical traffic and security monitoring systems," 2014, pp. 225–226.

[10] S. Wang and W. Kelly, "invideo - a novel big data analytics tool for video data analytics," 2014, pp. 1–19.

[11] S. Wang and R. Ledley, "Modified neumann architecture with micro-os for security," 2007, pp. 303–310.

[12] S. Wang, F. Shao, and R. S. Ledley, "Connputer - A framework of intrusion-free secure computer architecture," in *Proceedings of the 2006 International Conference on Security & Management, SAM 2006, Las Vegas, Nevada, USA, June 26-29, 2006*, 2006, pp. 220–225.

[13] Á. Serrano-Laguna, J. Torrente, P. Moreno-Ger, and B. Fernández-Manjón, "Tracing a little for big improvements: Application of learning analytics and videogames for student assessment," in *Fourth International Conference on Games and Virtual Worlds for Serious Applications, VS-GAMES 2012, Genoa, Italy, October 29-31, 2012*, 2012, pp. 203–209. [Online]. Available: http://dx.doi.org/10.1016/j.procs.2012.10.072

[14] V. SRideout, U. Foehr, and D. Roberts. (2010) Generation m2: Media in the lives of 8- to 18-year-olds. [Online]. Available: http://kaiserfamilyfoundation.files.wordpress.com/2013/01/8010.pdf

[15] H. Choi and S. Johnson, "The effect of context-based video instruction on learning and motivation in online courses," *The American Journal of Distance Education*, vol. 19, no. 4, pp. 215–217, 2005.

[16] S. Wang and M. Behrmann. (2010) Video indexing and automatic transcript creation.

[17] T. Elias. (2011) Learning analytics: Definitions, processes and potential. [Online]. Available: http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf

[18] C. E. Hmelo-Silver, C. A. Maher, A. Alston, M. Palius, G. Agnew, R. Sigley, and C. M. Mills, "Building multimedia artifacts using a cyber-enabled video repository: The vmcanalytic," in *46th Hawaii International Conference on System Sciences, HICSS 2013, Wailea, HI, USA, January 7-10, 2013*, 2013, pp. 3078–3087. [Online]. Available: http://dx.doi.org/10.1109/HICSS.2013.122

[19] C. E. Hmelo-Silver, C. A. Maher, G. Agnew, M. Palius, and S. J. Derry, "The video mosaic: design and preliminary research," in *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences, ICLS '10, Chicago, IL, USA, June 29 - July 2, 2010, Volume 2*, 2010, pp. 425–426. [Online]. Available: http://dl.acm.org/citation.cfm?id=1854729

[20] G. Agnew, C. M. Mills, and C. A. Maher, "Vmcanalytic: Developing a collaborative video analysis tool for education faculty and practicing educators," in *43rd Hawaii International International Conference on Systems Science (HICSS-43 2010), Proceedings, 5-8 January 2010, Koloa, Kauai, HI, USA*, 2010, pp. 1–10. [Online]. Available: http://dx.doi.org/10.1109/HICSS.2010.438

[21] C. A. F. P. Filho, T. A. Buck, and C. A. S. Santos, "An environment for video content indexing and retrieval base don visual features," in *Proceedings of the XV Brazilian Symposium on Multimedia and the Web*, ser. WebMedia '09. New York, NY, USA: ACM, 2009, pp. 25:1–25:8. [Online]. Available: http://doi.acm.org/10.1145/1858477.1858502

[22] D. M. Norris and L. Baer, "Building organizational capacity for analytics: Panel proposal," in *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, ser. LAK '12. New York, NY, USA: ACM, 2012, pp. 18–19. [Online]. Available: http://doi.acm.org/10.1145/2330601.2330612

[23] A. L. Dyckhoff, D. Zielke, M. Bültmann, M. A. Chatti, and U. Schroeder, "Design and implementation of a learning analytics toolkit for teachers," *Educational Technology & Society*, vol. 15, no. 3, pp. 58–76, 2012. [Online]. Available: http://www.ifets.info/download_pdf.php?j_id=56&a_id=1257

[24] S. Wang, J. Chen, and M. Behrmann. (2004) Visualizing search engine results of data-driven web content. [Online]. Available: http://www.w3.org/WAI/RD/2003/12/Visualization/VisSearch/VisualSearchEngine.htm

[25] S. Wang and M. Behrmann, "Automatic adaptive assessment in mlearning," 2009, pp. 435–438.

 ISBN: 978-1-61208-559-3