Internet Research Support System: An Application for Immediate Feedback in Authentic Internet Research

Nasiroh Omar
School of Computer Science and Information Technology, University of Nottingham, UK
nxo@cs.nott.ac.uk

Colin Higgins[1], Colin Harrison[2] and Diego Campo Millan[1]
[1]School of Computer Science and Information Technology and [2]School of Education, University of Nottingham, UK
cah@cs.nott.ac.uk,Colin.Harrison@nottingham.ac.uk

Abstract – **A prototype Internet Research Support System is presented based on earlier work on the modeling of online reading behavior and evaluating the reading process and its outcomes. Users complete a reading task by browsing the Web and writing a short essay. The online browsing data is captured in real-time, which, along with the essay, reflects the participants' online reading process. Verbatim Quotient Detection is possible which monitors the proportion of material copied from the Internet. A Latent Semantic Analysis component to evaluate sentence-to-sentence cohesion in the final essay is also described, and is in the process of being integrated with the main system. The system evaluation capabilities provide the participants with immediate feedback that is specific to their final essay in order to help them increase their online reading comprehension. It also archives data for research and evaluation purposes. One hundred and seventeen undergraduate students and one faculty member participated in two experiments to test the system where the outcomes were marked by the system and human markers. Analysis of the online reading activity demonstrated a negative correlation between the human marker's score and the number of verbatim words in the final essay. Furthermore, the analysis showed that when the whole Internet was available, even with limited time and a common task, there was significant variation in participants' search strategies, search terms and research processes.**

Index Terms – **Internet research, online reading comprehension, verbatim copied, Latent Semantic Analysis.**

Based on "Evaluating Real-time Online Research Data (RORD) and Verbatim Quotient Detection (VQD): Low Inference Tools to Monitor Outcomes of Unconstrained Authentic Internet Research", by Omar, N., C. Higgins, C. Harrison, and D. Campo Millan which appeared in the Proceedings of IEEE International Conference on July 5-8 2005, Kerkrade, Netherland, July 2006. © 2006 IEEE.

I. INTRODUCTION

This paper describes work that builds upon an earlier experiment [1] for testing an extended prototype version of an Internet Research Support system. It is based on previous work by the authors in the fields of Internet research, reading comprehension and automatic feedback.

In order to provide an online reading support tool, we need first to develop tools to evaluate online reading performance. Current computer-based online reading comprehension assessment is mostly based on multiple-choice type of questions [20], non multiple-choice questions and short essays [21-23] and comprehension summary [23]. However, Harrison [24] suggests that the most useful reading assessment is based on performance with authentic tasks. Such tasks may be devised using digital text reading due to the capability of the computer to capture the process of reading in digital contexts, and especially in web reading [25]. The focus of this paper is to further analyze and understand the general behaviour of online readers using the enhanced prototype Internet Research Support System that incorporates authentic immediate online reading feedback.

Performing Internet research requires individuals to have various literacy skills especially reading (while evaluating search results and comprehending web pages) and writing (while expressing reading comprehension in written form) [13]. Many studies analysing online reader behaviour have reported that some online readers encounter reading or navigation problems while performing internet research and are in need of various types of support [14]. It is reported that different online readers carry out Internet research in different styles and the research styles while working on goal-directed tasks display distinctive navigation characteristics [15]. McEneaney argued that navigational competence makes a significant contribution to online reading effectiveness. If

this shortcoming were left unsolved, the second-level digital divide in which users who have the ability to access online information but who will fail to comprehend the content [16] will become a serious problem. Therefore, with the unprecedented pace of digital text spreading [17], the medium of reading is changing [18, 19], and the need to have access to online reading support tools is increased.

Publications defining[2], modeling[3], evaluating [4, 5] and supporting Internet research [6-9] work are widespread, as are studies in the fields of reading comprehension [10] and automatic feedback [11, 12]. It is not the intention of this paper to review or summarize work in the three fields but to focus upon the role of automatic feedback in online reading comprehension in order to provide support while performing Internet research task.

This section begins with a brief introduction to the Internet research problem. It documents how Internet research activities relate to online reading comprehension and what attributes of online reading comprehension will be used as the basis for automating immediate feedback of the Internet research performance.

II. Understanding what online readers do while reading online

We began our experimental work by collecting data on the complexities of online reader behaviour. The term 'online reading behaviour' refers to the set of observable actions readers undertake while completing an online reading task. In the online environment we constructed, a research task is presented on a tabbed-pane in the system. Upon reading the online task, a reader will select a navigation goal to obtain a set of navigation options. The navigation goal is defined by the keyword(s) personally chosen by the reader and presented to a web search engine. The search engine will then list a sequence of relevant web sites. The sequence of web sites listed by the search engine for the reader to visit defines the "navigation options". 'Reading' is defined as a condition whereby the computer is asked to retrieve a web page by the reader. The machine will automatically retrieve the page via the web address when the reader clicks on it. The reading event occurs when a reader is visiting a web page.

We have modeled the behaviour of online reading (see [26] for further details). The aim of this model was to define and enable monitoring of the observable processes and output of online reading which could be captured physically by the computer. Once the model has been established we can define what is meant by online reading and therefore classify the processes and output to look for in evaluating online reading. We have designed and implemented a prototype system that can capture the low inference aspects of online reading processes and output. The processes and output of online reading are expressed in terms of user actions and user written output. This type of assessment belongs to the 'free response' type of assessment. The system is initially implemented as stand-alone. However, since the system involves online reading and capturing online reading activities, it is able to communicate with the web browser software. Given the lack of suitable tools for collecting online real time data that works for all web browsers, the prototype system will allow the online reader to use only one web browser - Internet Explorer, however, the online reader is free to visit any web sites available online (in what we term an unconstrained Internet research environment). In this paper we focus on monitoring user research, with particular emphasis on analyzing the verbatim text copied from the visited web sites and on the search terms used to navigate to this material.

III. The Internet Research Support System

The main focus of this research is to investigate a model of online reading comprehension behavior in an Internet research setting; and to analyze, design, implement and evaluate the development of an Internet Research Support System which supports online reading comprehension in open-ended, free-response, Internet research, by providing instant formative feedback. In its analysis phase, three major components have been identified. The three components are designed in order to accomplish the following aims:

1. Integrating related online reading comprehension activities on a single platform.
2. Embedding real time data collection mechanisms into the proposed system.
3. Integrating evaluation mechanisms into the online reading comprehension environment that facilitates realistic experimentation and evaluation of free-response online reading comprehension evaluation (RCE).

Hence, this prototype has the ability to allow its users to perform Internet researching and online reading, to collect real-time relevant data in terms of the online reading process and written output, to process the captured data automatically and to provide immediate feedback in relation to the amount of verbatim material copied.

Basically the Internet Research Support System becomes an interface between the online reader, the browser and the student's written output.

The prototype of the Internet Research Support System consists of online reading accessories and two components of online reading evaluation. The online reading accessories have three interchangeable parts: access to the chosen browser (Internet Explorer) for navigation purposes; a text pane from which to read the online reading task and a text pane into which the student can paste or write his/her own words found in the visited web pages. The online reader simply matches the commands and events provided by the system to activate either one.

The nature of online reading comprehension complexities allows varieties of online reading

comprehension evaluation mechanisms to be investigated, specified, designed and implemented. However, only two components of the evaluation elements (Latent Semantic Analysis (LSA) and Verbatim Quotient Detection (VDQ)) have been integrated into the prototype system.

LSA is a technique using a vector-based approach to automatically construct a semantic representation of a unit of text in large corpora where the representation is dependent on the corpora [27, 28]. LSA has been widely used in the evaluation of written answers [29-31] in examination or testing contexts, and the team at Nottingham has developed and is evaluating its own online LSA-based tools and software.

Nonetheless, our focus in this paper is on testing the prototype's Verbatim Detection Quotient and its immediate feedback tools. The VDQ has been designed with the following intentions:

1. The component should provide a mechanism to extract the URLs of the visited Internet resources from the low inference online reading comprehension activities, captured during the online reading comprehension session.
2. The component should be able to visit each of the extracted URLs, retrieve its content, compare it with the online reader's written text and store all information regarding verbatim copying in terms of the list and count of words copied and their URLs. The system should also calculate and store the overall percentage of the verbatim copy found in the online reader's written text.

The VDQ has been fully implemented and performs detection of the verbatim text copied from web pages visited by the online reader, and returns a figure in the range 0-100 indicating the percentage of text in the writing pane that is derived verbatim from the web source material accessed in that session. We define "verbatim copied" as any contiguous string of five or more words found in the text pane file. This percentage figure becomes the basis of the verification process.

By activating the Internet Research Support System, the reader has immediate access to the reading task pane, a web browser to perform online reading activities and to the text pane for writing/copying. The system records the activity start and end times. After conducting some web searches and after having either written some text or copied some text into the text pane, the online reader is able to activate the verbatim detection. The verbatim detection tool converts the Internet Explorer history file into an appropriate format [32], which is then used by the Real-time Online Research Data component to check the time stamped files. Only those URLs visited within the time parameters will be considered for verbatim detection. The verbatim detection tool accesses the web pages and compares the content with the words in the written text. If there are five or more contiguous words copied then they will be flagged as such.

At the end of the process a figure will be produced to indicate what percentages of words in the final text-pane document have been copied verbatim from Internet sources. Figure 1 illustrates the architecture of the Internet Research Support System.

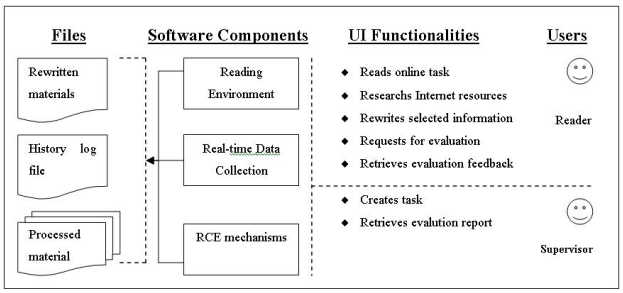

Figure 1: Part of the architecture of Internet Research Support System (formerly referred as Intelligent Online Reading Assessment System).

IV. METHODOLOGY

A. Experiments

In this study we were interested in evaluating online readers' data, with particular emphasis on the amount of verbatim material in their output file, which was captured using the Real-time Online Research Data capture component embedded in our Internet Research Support System. Two experiments have been conducted, one in 2005 (with 57 participants using the first version of the prototype) and one in 2006 (with 60 participants using the enhanced version of the prototype). The participants were first-year undergraduate students taking a Communication Skills in Computer Science module at the University of Nottingham, England, and one faculty member, who participated in the first experiment. All participants were given one hour to complete a research task, using web resources to answer a question, in not more than 500 words, of which up to 300 (60%) could be verbatim quotation. The work they undertook in this task was authentic in that it was to be integrated into a subsequent piece of coursework for the module, where the mark they recieved for the online reading activity contributed 10% to their overall module mark. The search task was defined as follows:

> "Shannon's first theorem in Information Theory enables us to calculate the entropy of a system of events if we know the probability of each individual event occurring. We can then use the Huffman algorithm to construct a compact code for the entire system. Recall that in general any coding of a system will not achieve the entropy limit. That is to say it will not generally be 100% efficient -- there will almost certainly be some redundancy.
>
> Information can be compressed if a new encoding can be found that reduces redundancy. If the original information can be totally recovered by undoing this encoding (i.e. by *decoding*) then the compression is called *lossless*; if so much redundancy has been removed that the original information cannot be fully recovered on decoding, then the compression is called *lossy*.

Making careful use of sources on the Internet, write an essay about lossless and lossy encoding of information, with particular respect to LZW encoding of text, JPEG encoding for still pictures and MPEG encoding for DVDs. The essay should be no more than 500 words long of which at least 40% should be your own words."

The experiment took place in a network lab. Even though the prototype of the IRSS is designed to be a stand-alone system, the system is capable of accessing the web browser. Every participant used the same hardware and software. Even though there are some slight differences in the system user interface between the first experiment and the second experiment, the system functions were the same. When the experiment took place, the participant logged on to his/her account and started the Real time Online Research Data capture component. The system saved their starting and ending times to be used in processing the real time collected data. At the end of the experiment the system converted the Internet Explorer history file into a readable file format (contained visited URLs) as input to the Verbatim Quotient Detection component.

B. Data

All the data were captured by using the data capture system embedded in the Internet Research Support System. The captured data for each participant were in the form of a text file consisting of all the visited URLs and a document file consisting of an essay.

C. Results

The Internet Research Support System delivers the output (shown in Figures 2 to 5) for each participant. The system displays eight step-by-step instructions at the top of its main window. Figure 2 shows the participant's starting page where he/she needs to fill in his/her username, to read the given task displayed on the 'Read Task' text panel and to click the 'Start' button in order to start the online reading session. The system needs the user identification in order to create a directory so that all relevant files produced during the online reading session can be retained. The user is then expected to click the 'Launch' button to activate the Internet Explorer browser. Figure 3 shows the text panel in which the user writes his/her online reading summary. When the user clicks the 'Evaluate' button, a feedback page including the participant's written text is displayed on the 'Read Evaluation' text panel. At the bottom of the output window, the number of participant's written words, the number of verbatim words, the percentage of verbatim words and the visited web pages which have participant's verbatim material are shown in the right side of the IRSS window.

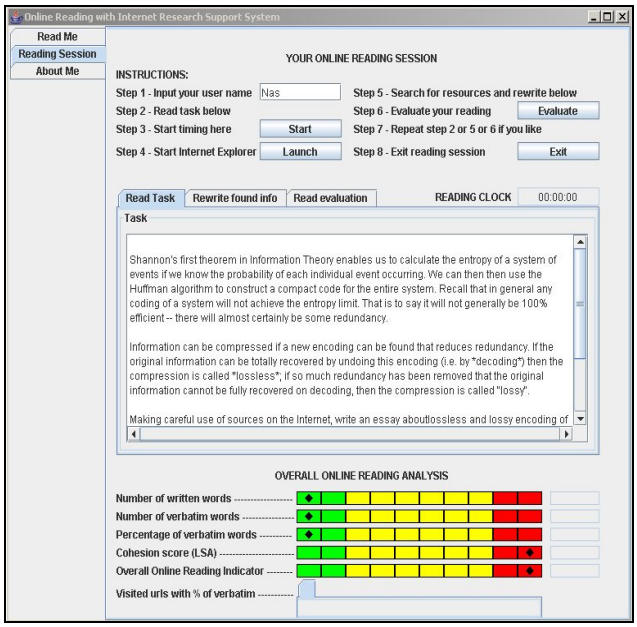

Figure 2: The sample output displaying the starting point of an online reading session.

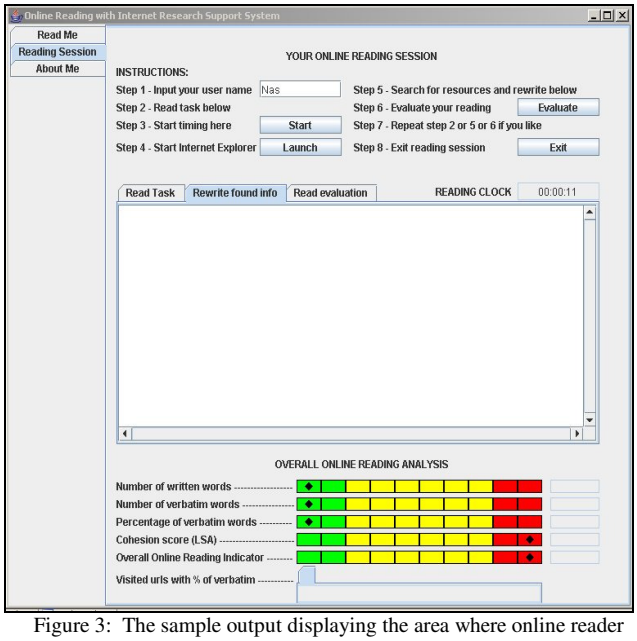

Figure 3: The sample output displaying the area where online reader is expected to write online reading summary.

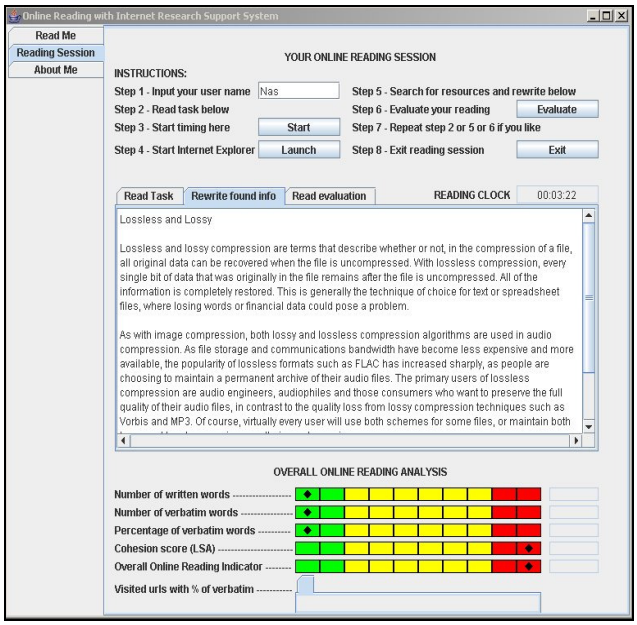

Figure 4: The sample output displaying online reading summary.

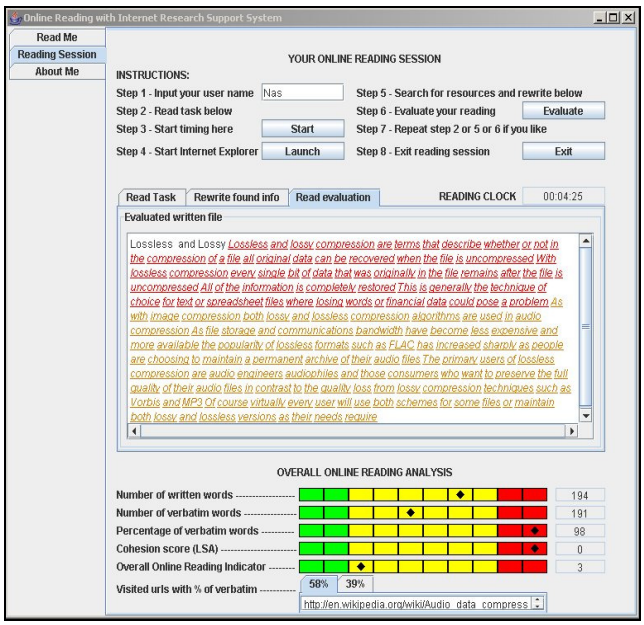

Figure 5: The sample output displaying the online reading feedback.

All the results were transferred to a Microsoft Excel file. Subsequently, four sets of analyses were carried out. The first set analysed the number of verbatim words present in the final output files of the online reader. The second calculated the percentage of verbatim words out of the overall number of words produced by the online readers. The third and fourth sets were conducted to evaluate how the online readers behaved when they selected the URLs and search terms.

In Experiment 1, only 10 out of 58 participants copied more from the web than the 60% they were allowed. However the percentage of verbatim text copied was great, ranging from 0% to 95%. In Experiment 2, the results confirmed those in the previous experiment. In Experiment 2, only 4 out of 60 participants copied from the web more than they were allowed. The percentage of verbatim text copied ranged from 0% to 84%.

A low or 0% verbatim score does not necessarily imply that the final text was wholly different to the original Internet source. The 0% score might have resulted from the reader rephrasing the sentences so that the number of running words copied was less than five. In Experiment 1, participants did not have verbatim score data available to them; however, in Experiment 2, verbatim text was both presented and highlighted in the text pane, although the five-words algorithm was not explained. Table 1 shows the highest, lowest, mean and standard deviation of material copied for both of the experiments.

Table 1: The highest, lowest, average and standard deviation VDQ scores for all the participants (n= 58)

Expr	Highest Percentage	Lowest Percentage	Average Percentage	Standar d Deviation
1	95.00	0.00	33.34	24.90
2	81.00	0.00	34.20	22.30

In Experiment 1, ten participants exceeded the 60% maximum verbatim allowed while 26% of all participants copied verbatim more than they wrote using their own words. However, in Experiment 2, where the percentage of verbatim words copied was made available to the participants in the online reading session, the percentage of participants exceeding the maximum verbatim allowed was reduced to 7%.

We investigated whether there was a correlation between a high verbatim score and low essay score. A statistical analysis of the relationship between the participants' percentage of verbatim copied and human marker's mark in Experiment 1 shows a slight but statistically significant negative correlation of -0.24 where $n = 58$ and $p < 0.01$.

Figures 6, 7, 8 and 9 demonstrate that students' research strategies show a high degree of individuality in both experiments. In Experiment 1 79% of all the total of 600 URLs visited by the participants were visited only once and 77% of the search phrases used in navigating to these sites were unique, and based on word combinations that were not used by others. In Experiment 2, 73% of 483 URLs were visited only once by the participants and 84% of the search phrases used in their navigation were unique.

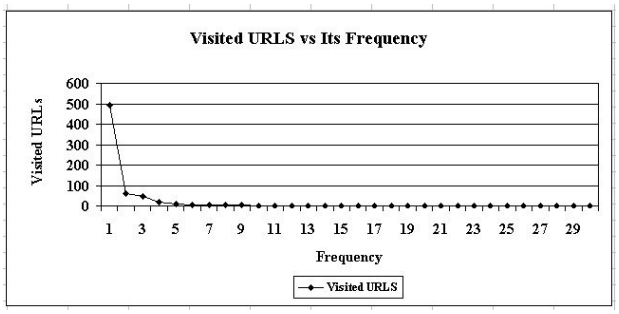

Figure 6: Shows the number of unique URLs visited by the participants in the study, with their frequency in Experiment 1.

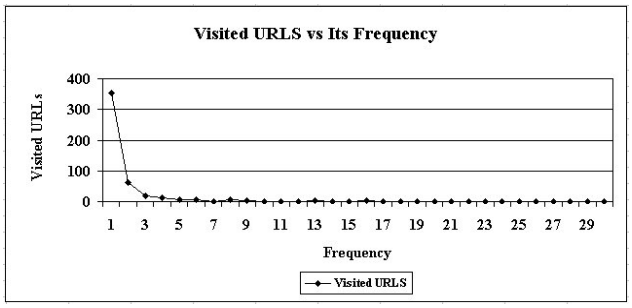

Figure 7: Shows the number of unique URLs visited by the participants in the study, with their frequency in Experiment 2.

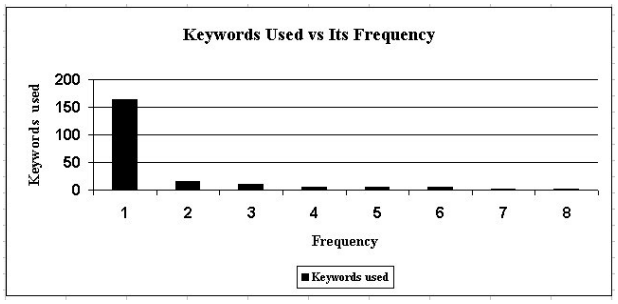

Figure 8: Shows the number and frequency of different Google keywords used in navigation (ie most words were used once only; fewer than 20 were used by two participants, etc.) in Experiment 1.

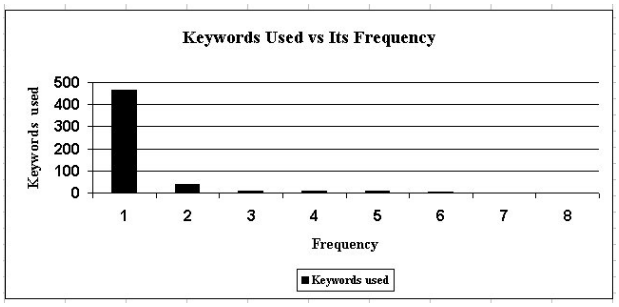

Figure 9: Shows the number and frequency of different Google keywords used in navigation (ie most words were used once only; fewer than 20 were used by two participants, etc.) in Experiment 2.

We interpret these results as indicating that Internet research strategies are much more varied than is generally supposed. This wide variation in search terms and navigation end points is also perhaps an indication that the research strategies adopted in this study were authentic.

V. CONCLUSION

Our current research is a continuation of the work presented in [26] and [1]. The work is still in progress to establish the extent to which online reading processes and output can be evaluated. We have improved our approach by collecting our data automatically in ways recommended by McEneany in [33] and included a verbatim text detection tool which we made more transparently available to participants during the Real-time Online Research Data process in the second experiment. In this paper we have outlined some aspects of the search and composition behaviour of online readers from which we aim to construct an indication of how well they read in terms of online processes based on the verbatim aspect. The results in the second experiment confirmed that the online readers' processes and products were highly authentic which we suggest was also the case in the first experiment. In the future we will report on further analyses, based on more complete models of expert reader behaviour, and, where appropriate, incorporate online Latent Semantic Analysis sentence-to-sentence cohesion data derived from matching user navigation and text pane data against our own corpora.

REFERENCES

[1] Omar, N., C. Higgins, C. Harrison, and D. Campo Millan. *Evaluating Real-time Online Research Data (RORD) and Verbatim Quotient Detection (VQD): Low Inference Tools to Monitor Outcomes of Unconstrained Authentic Internet Research* in *6th. IEEE International Conference on Advanced Learning Technologies* 2006: IEEE Computer Society.

[2] Broder, A., *A Taxanomy of Web Search.* SIGIR Forum, 2002. **36**(2).

[3] Juvina, I., H. van Oostendorp, P. Karbor, and B. Pauw. *Towards Modeling Contextual Information in Web Navigation*. in *27th. Annual Meeting of the Cognitive Science Society*. 2005: The Cognitive Science Society, Inc.

[4] Bruza, P., R. McArthur, and S. Dennis. *Interactive Internet Search: Keyword, Directory and Query Reformulation Mechanisms Compared*. in *23rd. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2000. Athens, Greece.

[5] Rose, D.E. and D. Levinson. *Understanding User Goals in Web Search*. in *WWW 2004*. 2004. New York, USA.

[6] Lee, U., Z. Liu, and J. Cho. *Automatic Identification of User Goals in Web Search*. in *International World Wide Web Conference (WWW 2005)*. 2005. Chiba, Japan.

[7] Olston, C. and E.H. Chi., *ScentTrails: Integrating Browsing and Searching on the Web.* ACM Transactions on Computer-Human Interaction, 2003. **10**(3): p. 177-197.

[8] Teevan, J., C. Alvarado, M.S. Ackerman, and D.R. Karger. *The Perfect Search Engine Is Not Enough: A Study of Orienteering Behavior in Directed Search*. in *Computer Human Interaction*. 2004. Vienna, Austria.

[9] Aula, A., N. Jhaveri, and M. Käki. *Information Search and Re-access Strategies of Experienced Web Users*. in *WWW 2005*. 2005. Chiba, Japan: ACM.

[10] Harrison, C., *Understanding Reading Development*. 2004, London: SAGE.

[11] Tsintsifas, A., *A Framework for the Computer-Based Assessment of Diagram-Based Coursework*, in *Computer Science and Information Technology*. 2002, University of Nottingham: Nottingham.

[12] Higgins, C.A. and B. Bligh. *Formative Computer Based Assessment in Diagram Based Domains*. in *11th. Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiCSE 06)*. 2006. Bologna, Italy.

[13] Henry, L., *Information Search Strategies on the Internet: A Critical Component of New Literacies.* Webology, 2005. **2**(1).

[14] McEneaney, J.E. *Navigational Correlates of Comprehension in Hypertext*. in *11th. Hypertext Conference of the Association for Computing Machinery*. 2000. San Antonio, Texas: New York, NY: ACM Press.

[15] Juvina, I. and H. van Oostendorp. *Predicting User Preferences - from Semantic to Pragmatic Metrics of Web Navigation Behaviour*. in *Ducth Directions in HCI*. 2004. Amsterdam.

[16] Hargittai, E., *Beyond Logs and Surveys: In-Depth Measures of Peoples Web Use Skills.* Journal of American Society for Information Science and Technology, 2002. **53**(14): p. 1239-1244.

[17] Sutherland-Smith, W., *Weaving the Web literacy: changes in reading from page to screen.* Reading Teacher, 2002. **55**(7): p. 662-671.

[18] Coiro, J., *Reading Comprehension on the Internet: Expanding Our Understanding of Reading Comprehension to Encompass New Literacies.* The Reading Teacher, 2003. **56**(6).

[19] Johnson-Eilola, J. and A.C. Kimme Hea, *After Hypertext: Other Ideas.* Computers and Composition, 2003. **20**: p. 415-425.

[20] IRA, *High Stakes Assessments in Reading: A Position Statement of the International Reading Association.* Reading Teacher, 1999. **53**(3): p. 257-263.

[21] Burstein, J., C. Leacock, and R. Swartz. *Automated Evaluation Essays and Short Answers*. in *Computer Aided Assessment*. 2001. Loughborough, UK.

[22] Mason, O. and I. Grove Stephensen. *Automated Free Text Marking with Paperless School*. in *Computer Aided Assessment*. 2002. Loughborough, UK.

[23] Hepplestone, S. and P. Helm. *Strategies for Coping with Unexpected Uptake of CBA*. in *Computer Aided Assessment*. 2003. Loughborough, UK.

[24] Harrison, C., N. Omar, and C. Higgins, *Postmodern Principles for Responsive Reading Assessment: a Case Study of a Complex Online Reading Task* in *Assessing Reading: From Theories to Classrooms*, M. Sainsbury, C. Harrison, and A. Watts, Editors. 2006, National Foundation for Educational Research.: Slough.

[25] Hawkins, J., J. Frederiksen, A. Collins, B. Dorothy, and E. Collins, *Assessment and Technology.* Communications of ACM, 1993. **36**(5).

[26] Omar, N., C. Higgins, and C. Harrison. *Investigating an Approach for Online Reading Assessment*. in *5th. IEEE International Conference on Advanced Learning Technologies, ICALT 2005* 2005: Nottingham, U.K.

[27] Schunn, C.D. *The Presence and Absence of Category Knowledge in LSA*. in *21st. Annual Conference of the Cognitive Science Society*. 1999: Mahwah, NJ:Erlbaum.

[28] Landauer, T.K., P.W. Foltz, and D. Laham, *An Introduction to Latent Semantic Analysis.* Discourse Processes, 1998. **25**(2/3): p. 259-284.

[29] Kintsch, E., D. Steinhart, G. Stahl, C. Matthews, and R. Lamb, *Developing Summarization Skills Through the Use of LSA-Backed Feedback.* Interactive Learning Environments, 2000. **8**(2): p. 87-109.

[30] Lemaire, B. and P. Dessus, *A system to Assess the Semantic Content of Student Essays.* Journal of Educational Computing Research, 2001. **24**(3): p. 305-320.

[31] Kanijeya, D., A. Kumar, and S. Prasad. *Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA*. in *HLT-NAACL 2003*. 2003.

[32] NirSofer, *Iehv*. 2006, NirSoft Freeware.

[33] McEneaney, J.E. *New Approaches to Data Collection and Analysis in Online Reading Studies*. in *Annual Meeting of the National Reading Conference*. 1999. Orlando, FL USA.

Nasiroh Omar is a PhD research student in Learning Technology Research Group in the University of Nottingham.

Colin Higgins is an Associate Professor in the School of Computer Science in the University of Nottingham. He is head of the Learning Technology Research Group where his research focuses on automatic assessment (particularly of free response answers), the teaching of programming and pattern recognition with emphasis on handwriting recognition.

Colin Harrison is Professor of Literacy Studies in Education in the University of Nottingham. He is a former editor of the Journal of Research in Reading, and past President of the United Kingdom Reading Association. His current research work focuses on evaluating the uses of new technologies in school and non-school settings, and on reading and comprehension processes; Interactive Classroom Explorer, the interface he developed for mediating online video for teacher development via a bulletin board, is currently being evaluated in three national projects in the UK.