Preview of Award 1319578 - Annual Project Report

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	1319578
Project Title:	III: Small: Integrated Digital Event Archiving and Library (IDEAL)
PD/PI Name:	Edward A Fox, Principal Investigator Kristine Hanna, Co-Principal Investigator Andrea L Kavanaugh, Co-Principal Investigator Steven D Sheetz, Co-Principal Investigator Donald J Shoemaker, Co-Principal Investigator
Recipient Organization:	Virginia Polytechnic Institute and State University
Project/Grant Period:	09/01/2013 - 08/31/2016
Reporting Period:	09/01/2013 - 08/31/2014
Submitting Official (if other than PD\PI):	N/A
Submission Date:	N/A
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	N/A

Accomplishments

*** What are the major goals of the project?**

We will ingest tweets and Web-based content from social media and the general Web, including news and governmental information. In addition to archiving materials found, we will build an information system that includes related metadata and knowledge bases, consistent with the 5S (Societies, Scenarios, Spaces, Structures, Streams) framework, along with results from our intelligent focused crawler, to support comprehensive access to event related content. With the support of key partners, the IDEAL team will undertake important research, education, and dissemination efforts, to achieve three complementary objectives:

1. Collecting: We will spot, identify, and make sense of interesting events. We also will accept specific or general requests about types of events. Given resource and sampling constraints, we will integrate methods to identify appropriate URLs as seeds, and specify when to start crawling and when to stop, with regard to each event or sub-event. We will integrate focused crawling and filtering approaches in order to ingest content and generate new collections, with high precision and recall.

2. Archiving & Accessing: Permanent archiving, and access to those archives, will be ensured by our partner, Internet Archive (IA). Immediate access to ingested content will be facilitated through big data software built on top of our new Hadoop cluster.

3. Analyzing & Visualizing: We will provide a wide range of integrated services beyond the usual (faceted) browsing and searching, including: classification, clustering, summarization, text mining, theme and topic identification, and visualization.

*** What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?**

Major Activities:

We have collected tweets and Web collections about many events. Web collections were archived using IA software. Tweet collections were stored on local servers.

Collections have been indexed and made available for searching, browsing, and other services.

We developed and applied software and tools for collecting, storing, organizing, indexing, and analyzing Web and tweet collections. Interfaces and visualization methods were developed for accessing and visualizing collections. One dissertation and reports from projects in three courses have been prepared, and proposals for two closely related dissertations have been approved.

We started collaborations with universities in Egypt, Tunisia, Mexico, and the Philippines. Together we designed, conducted, and analyzed survey studies in Egypt, Tunisia, and Mexico -- supporting the validation and analysis of our collections.

Specific Objectives:

Web and tweet collections were built using IA, focused crawling, and tweet archiving tools. Collections span many kinds of events and places around the world, covering the previous year.

A prototype spreadsheet-style interface was developed for accessing and analyzing different collections, through one of multiple class projects.

	Topic identification studies were made on several collections, leading to a dissertation and then a class project using Wikipedia content.
Significant Results:	We have built Web collections about the Washington State mudslide, Kenya mall attack, Typhoon Haiyan, and a New York building collapse, using an event focused crawler. We have built tweet collections about: California shooting (May 2014), chemical spill in West Viriginia (Jan. 2014), Colorado flood (Sept. 2013), Hurricane Arthur (July 2014), Kenya mall attack (Sept. 2013), Manhattan building explosion (Mar. 2014), Malaysia Airlines plane crash (Mar. 2014), Mexico earthquake (Apr. 2014), Navy Yard shooting (Sept. 2013), Nevada school shooting (Oct. 2013), New Mexico school shooting (Jan. 2014), Nigeria school attack (Feb. 2014), Pennsylvania school stabbing (Apr. 2014), South Korea Ferry Sank (Apr. 2014), Turkey mine accident (May 2014), Typhoon Yolanda (Nov. 2013), UK floods (Feb. 2014), Ukraine protest (Nov. 2013), US midwest tornadoes (Nov. 2013), Washington landslide (Mar. 2014), wildfires in San Diego (May 2014), and Winter Storm Pax (Feb. 2014). Many other collections were built too, including about accidents, community activities, disease outbreaks, politics, revolution, and specific organizations of interest. Web collections were prepared and made accessible through a SOLR REST interface for topic identification, classification, clustering, and summarization. Web collections are available through the project website. A Hadoop cluster was constructed from parts by interested students to support this project, software was set up, and students in classes as well as other volunteers have helped make it useful for IDEAL.
Key outcomes or Other achievements:	Student project reports have documented the learning and findings of the students already involved, and are available online for others to learn from too (see below under Other Publications). Other publications and presentations have helped disseminate results and expand our collaboration with partners and stakeholders.

*** What opportunities for training and professional development has the project provided?**

Research has involved IDEAL staff and other students in developing an event focused crawler prototype and our new Hadoop cluster.

Two courses (undergraduate and graduate) were given that included class presentations and term projects related to the IDEAL project. Three students in an undergraduate research course have focused on helping with IDEAL.

Connections were established with the University of the Phillipines, Diliman, for studies about Typhoon Yolanda

Collaborators from Tunisia, Egypt, and Mexico have been gaining experience about extending their research through IDEAL, learning more about digital libraries and archiving.

*** How have the results been disseminated to communities of interest?**

The IDEAL project website has provided information and pointers for previous and current Web collections through IA and summaries for tweet collections.

PI Edward Fox attended the NSF WIRE workshop (http://wp.comminfo.rutgers.edu/nsfia/), supported also by the Internet Archive, that discussed the future of Web archiving and research opportunites, presenting about IDEAL (see http://fox.cs.vt.edu/talks/2014/20140618FoxNSF-IA-WIRE.pptx).

2 ISCRAM posters (one connected with graduate research assistant Mohamed Farag attending the doctoral consortium) and Digital Government conference paper (with Mexico collaboration).

Co-PI Kavanaugh led an effort to analyze an archive of community/government content created as part of a separate but related NSF funded project, called "Participation on the Town Square in the Era of Web 2.0" (VTS) and has connected these two projects. In the VTS project, the researchers (Kavanaugh, PI) created an aggregated website of local community content, called the Virtual Town Square (http://vts.cs.vt.edu). Students in PI Fox's course analyzed the archive of aggregated RSS feeds from the Virtual Town Square (VTS), including local Web-based news and information, and user generated content, such as tweets, blogs, comments, and public group Facebook posts. The students worked closely with Drs. Fox and Kavanaugh and the graduate students involved in the VTS project to model topics across the collection that covers a time period of 8 months (September 2012 through October 2013). Modeling the topics over time in the geographic community can provide us with insights into the important community events during this period, and the local commentary around these events. It also reveals hidden (or less visible) topics and conversations that have occurred in the community.

*** What do you plan to do during the next reporting period to accomplish the goals?**

More webpages and tweets will be collected covering all relevant kinds of events.

We will continue to design, develop, refine, and test tools and other software allowing us and collaborators to analyze, visualize, and otherwise research with our event-related content.

The PI will teach a new course, Computational Linguistics, in the fall, and two classes next spring, in which students will work with IDEAL collections and tools. This fall the IDEAL collections will be the focus, as students learn about text analysis, extraction, and summarization. In the spring there should be term projects related to IDEAL.

The two graduate assistants, Mohamed Farag and Sunshin Lee, now that their dissertation proposals related to IDEAL have been approved, will work on their dissertations, as well as other aspects of IDEAL.

We will follow up with our various collaborators on survey studies, and leverage that work to help us better understand and improve the related collections and tools.

We will extend the many contacts made over the past year, including with various groups in the Library and Digital Humanities at Virginia Tech (e.g., with the Peace Center), as well as at other sites, to support the diverse current and prospective stakeholders who can benefit from IDEAL.

Products

Books

Edward A. Fox and Jonathan P. Leidig, editors (2014). *Digital Library Applications: CBIR, Education, Social Networks, eScience/Simulation, and GIS* March, 175 pages, ISBN paperback, ebook 9781627050333, http://dx.doi.org/10.2200/S00565ED1V01Y201401ICR032. Morgan & Claypool Publishers. San Francisco. Status = PUBLISHED;

Acknowledgment of Federal Support = Yes ; Peer Reviewed = No ; ISBN: 9781627050326

Edward A. Fox and Ricardo da Silva Torres, editors (2014). *Digital Library Technologies: Complex Objects, Annotation, Ontologies, Classification, Extraction, and Security* Ebook 9781627050319, http://dx.doi.org/10.2200/S00566ED1V01Y201401ICR033. Morgan & Claypool Publishers. San Francisco. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No ; ISBN: 9781627050302

Book Chapters

Conference Papers and Presentations
Andrea Kavanaugh, Steven Sheetz, John Tedesco, Sandoval-Almazan Rodrigo, and Edward Fox (2014). *Media Use during Conflicts: Information Gratifications & Efficacy during 2012 Mexican Elections*. 15th ACM Annual International Conference on Digital Government Research (dg.o 2014). Aguascalientes City, Mexico. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Inventions

Journals

Licenses

Other Products

Other Publications
Lech, Adam; Pontani, Joseph; Bollinger, Matthew (2014). *DLRL Cluster*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/47945. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Bonnefond, Ward; Menzel, Chris; Morris, Zack; Patel, Suhas; Ritchie, Tyler; Tedesco, Marcus; Zheng, Franklin (2014). *Developing an improved focused crawler for the IDEAL project*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/47939. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Aly, Mustafa; Gulotta, Gasper (2014). *IDEAL Pages*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/47938. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Farghally, Mohammed; Elbery, Ahmed (2014). *IDEAL Pages*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/47952. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Burnett, Austin; Neuman, Shawn; Ardura, Anthony; Lacy, Rex (2014). *IDEAL Spreadsheet*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/47942. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Sheetz, Steven; Kavanaugh, Andrea; Fox, Edward; Elmongui, Hicham; Hassan, Riham; Yang, Seugwon; Magdy, Mohammed; Shoemaker, Donald (2014). *Information Uses and Gratifications in Crisis: Student Perceptions since the Egyptian Uprising*. Poster in Proceedings of the 11th International ISCRAM Conference, University Park, Pennsylvania, USA, May 18-21. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Mohamed M. G. Farag and Edward A. Fox (2014). *Intelligent Event Focused Crawling*. Poster in Proceedings of the 11th International ISCRAM Conference, University Park, Pennsylvania, USA, May 18-21. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Roble, Benjamin; Cheng, Justin; Sbitani, Marwan (2014). *NRV Tweets and RSS feeds*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/47937. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Shuffett, Michael (2014). *Twitter Metadata*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/47949. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Joseph Sebastian (2014). *Undergraduate Research Final Paper (on Twitter data loading into IDEAL Hadoop cluster), http://eventsarchive.org/sites/default/files /SebastianCS2994ReportSpring2014.pdf*. Class short final report and presentation about assisting the Integrated Digital Event Archive (IDEAL) team in transferring their Twitter data on world events from a MySQL server into a Hadoop cluster. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Xuan, Zhang; Wei, Huang; Ji, Wang; Tianyu, Geng (2014). *Unsupervised Event Extraction from News and Twitter*. Technical Report, Virginia Tech, http://hdl.handle.net/10919 /47954. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Rinaldi, Anthony; Mehta, Dev (2014). *VT Web Archive Project*. Technical Report, Virginia Tech, http://hdl.handle.net/10919/47935. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Neidig, Sloane; Johnson, Samantha; Cabrera, David; Hoffman, Erika (2014). *Xpantrac Connection with IDEAL*. Technical Report, Virginia Tech, http://hdl.handle.net/10919 /47941. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Patents

Technologies or Techniques

Thesis/Dissertations
Seungwon Yang. *Automatic Identification of Topic Tags from Texts Based on Expansion-Extraction Approach, http://hdl.handle.net/10919/25111*. (2014). Virginia Polytechnic Institute and State University. Acknowledgement of Federal Support = Yes

Websites
Computation Linguistic Course
http://fox.cs.vt.edu/CS4984CL.htm

Homepage for fall 2014 class that will focus on the IDEAL collections and will aid work on IDEAL. This is supported in part through Villanova University (pass-through from NSF DUE-1141209): Computing in Context; for 8/15/2012 - 7/31/2015; sole PI (with overall PI at Villanova: Robert Beck)

Crisis, Tragedy, Recovery Network

http://www.ctrnet.net

Covering work on the prequel project, focused on various types of disasters, now incorporated into the IDEAL project website. From NSF IIS-0916733: III:Small:Integrated Digital Library Support for Crisis, Tragedy, and Recovery, 8/1/2009 - 7/31/2013, PI Edward A. Fox. Co-PIs: Naren Ramakrishnan, Steven Sheetz, Andrea Kavanaugh, Donald Shoemaker

Events Archiving
http://eventsarchive.org

Website for this project, IDEAL, including content from two prior NSF projects, IIS-0916733 and IIS-0736055

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Fox, Edward	PD/PI	2
Hanna, Kristine	Co PD/PI	0
Kavanaugh, Andrea	Co PD/PI	1
Sheetz, Steven	Co PD/PI	2
Shoemaker, Donald	Co PD/PI	1
Mansour, Riham	Other Professional	0
Chen, Yinlin	Graduate Student (research assistant)	2
Farag, Mohamed	Graduate Student (research assistant)	6
Lee, Sunshin	Graduate Student (research assistant)	6
Yang, Senugwon	Graduate Student (research assistant)	2
Ganotra, Ishita	Undergraduate Student	1
Jo, SoHyun	Undergraduate Student	1
Kaul, Rohan	Undergraduate Student	1
Kim, Jun	Undergraduate Student	1
Riblet, Stephen	Undergraduate Student	1
Sebastian, Joseph	Undergraduate Student	1
Kanan, Tarek	Other	1
Sun, Yue	Other	0

Full details of individuals who have worked on the project:

Edward A Fox
Email: fox@vt.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 2

Contribution to the Project: PI, directing all aspects of the project, giving presentations, preparing publications, preparing and teaching courses with related projects, supervising GRAs and student volunteers.

Funding Support: NSF

International Collaboration: Yes, Qatar
International Travel: Yes, Qatar - 0 years, 0 months, 14 days; - 0 years, 0 months, 0 days; - 0 years, 0 months, 0 days

Kristine Hanna
Email: kristine@archive.org
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 0

Contribution to the Project: Co-PI, responsible for Internet Archive aspects of project

Funding Support: Internet Archive

International Collaboration: No
International Travel: No

Andrea L Kavanaugh
Email: kavan@vt.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Co-PI Kavanaugh contributes to the social science, social computing, and HCI aspects of the project. Specifically, she has led the development of a study of users' experience of information seeking during and since the political crises of the Arab Spring (notably, Egypt and Tunisia). She worked closely with Prof. Steve Sheetz on designing and adapting for each country a survey questionnaire in collaboration with partner researchers in academic institutions in Tunisia and Egypt (and most recently, Mexico). Dr. Kavanaugh subsequently recruited and secured the research partners in Tunisia and Mexico. Our three partners administered the online survey to their university students. Dr. Kavanaugh and Prof. Sheetz worked closely with the collaborators on the analyses of the completed survey responses. Dr. Kavanaugh led two full research conference papers on the findings, presented at: 1) ISCRAM, and 2) the 15th International Annual Conference on Digital Government Research. The ISCRAM conference paper was invited and published in expanded form in the International Journal of ISCRAM (late 2013). The Digital Government research paper, whose proceedings are published by ACM Press and appear in the ACM Digital Library, was nominated for best paper.

Funding Support: NSF

International Collaboration: No
International Travel: No

Steven D Sheetz
Email: sheetz@vt.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 2

Contribution to the Project: Co-PI helping with the Mexico data analysis and paper, as well as Tunisia data analysis. He has been working on a Decision Science paper using Egypt Round 2 data (still in progress and nearing completion, but with many days of work already). He prepared and presented an ISCRAM 2014 poster. He has led efforts related to the ontology, and joined many meetings.

Funding Support: NSF

International Collaboration: No
International Travel: No

Donald J Shoemaker
Email: shoemake@vt.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Co-PI providing social science expertise to the development of the website and other IDEAL efforts. He helps connect with local advisers, such as the Center for the Study of Peace and Violence Prevention at Virginia Tech. Professor Shoemaker, during a three month visit, met with the Vice Chancellor for Research and Development of the University of the Philippines. He and other colleagues and administrators at the University of the Philippines are working on disaster research topics. They are interested in working with the IDEAL team on cooperative research opportunities. In addition, the University of the Philippines is developing an interdisciplinary course on the causes and effects of natural disasters, including community responses to disasters, such as Typhoon Haiyan (Yolanda) which hit the southeast part of the country in November 2013. They are interested in working with members of the IDEAL team on providing material for this course.

Funding Support: NSF

International Collaboration: Yes, Philippines
International Travel: Yes, Philippines - 0 years, 3 months, 0 days

---

Riham Hassan Abdel-Moneim Mansour
Email: rihamma@microsoft.com
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 0

Contribution to the Project: Helped in research, design, and analysis of the survey about social media use during the Egyptian uprising. Helped supervise the work of Mohamed Farag. Earlier, at the Arab Academy of Science and Technology in Cairo, and having visiting Virginia Tech a few years ago and worked with the NSF funded project team, she recruited another Egyptian collaborator (Prof. Hicham Elmongui, of the University of Alexandria).

Funding Support: N/A

International Collaboration: No
International Travel: No

Yinlin Chen
Email: ylchen@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 2

Contribution to the Project: Helped with developing user interfaces and the website for the project using Drupal, Solr, and automatic classification methods

Funding Support: NSF

International Collaboration: No
International Travel: No

Mohamed Farag
Email: mmagdy@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 6

Contribution to the Project: Developed prototypes that demonstrate research goals and ideas. Helped guide seven class projects related to IDEAL during the spring. Helped with webpage collections and their processing. Prepared his doctoral proposal, successfully defended that, and has been carrying out the planned research related to IDEAL.

Funding Support: NSF

International Collaboration: No
International Travel: No

Sunshin Lee
Email: sslee777@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 6

Contribution to the Project: Developed prototypes that demonstrate ideas and goals of the project. Led work on tweet collection. Guided one undergraduate research student and multiple undergraduate class projects. Led the planning, ordering, construction, software setup, and operation of the Hadoop cluster. Prepared his doctoral proposal, successfully defended that, and has been carrying out the planned research related to IDEAL.

Funding Support: NSF

International Collaboration: No
International Travel: No

Senugwon Yang
Email: seungwon@vt.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 2

Contribution to the Project: Helped developed topic extraction prototype, Xpantrac, which is the subject of his dissertation. Working with the PI, he submitted a journal paper on this for publication.

Funding Support: NSF

International Collaboration: No
International Travel: No

Ishita Ganotra
Email: ishitag@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Took Undergraduate Research course in spring focused on IDEAL, in particular on the ontology, schema to describe classes of events, and related terminology, helping prepare for the fall course on Computational Linguistics that will summarize event collections.

Funding Support: N/A

International Collaboration: No
International Travel: No

SoHyun Jo
Email: sohyun@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Volunteer, helping with project research

Funding Support: N/A

International Collaboration: No
International Travel: No

Rohan Kaul
Email: rohan@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Took Undergraduate Research course in spring focused on IDEAL, in particular on the ontology, schema to describe classes of events, and related terminology, helping prepare for the fall course on Computational Linguistics that will summarize event collections.

Funding Support: N/A

International Collaboration: No
International Travel: No

Jun Kim
Email: junk91@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Volunteered to build user interfaces for accessing Web archives

Funding Support: N/A

International Collaboration: No
International Travel: No

Stephen Riblet
Email: sriblet@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Volunteer, exploring event spotting

Funding Support: N/A

International Collaboration: No
International Travel: No

Joseph Braeden Sebastian
Email: jbraeden@vt.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 1

Contribution to the Project: Volunteered to upload web and tweet archives to hadoop cluster

Funding Support: N/A

International Collaboration: No
International Travel: No

Tarek Kanan
Email: tarekk@vt.edu
Most Senior Project Role: Other
Nearest Person Month Worked: 1

Contribution to the Project: Doctoral student, helping prepare for the fall Computational Linguistics course that will use IDEAL collections, and assisting with search and Arabic research, using SOLR and other tools

Funding Support: Qatar

International Collaboration: Yes, Qatar

International Travel: No

Yue Sun
Email: syue88@vt.edu
Most Senior Project Role: Other
Nearest Person Month Worked: 0

Contribution to the Project: Doctoral student, helping analyze survey data

Funding Support: N/A

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?

Name	Type of Partner Organization	Location
Alexandria University	Academic Institution	Egypt
Autonomous University of the State of Mexico, Toluca	Academic Institution	Toluca, Mexico
High Institute of Management of Tunis	Academic Institution	Tunis, Tunisia
Internet Archive	Other Nonprofits	San Francisco, CA, USA
University of the Philippines, Diliman	Academic Institution	Philippines

Full details of organizations that have been involved as partners:

Alexandria University

Organization Type: Academic Institution
Organization Location: Egypt

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: collaborated on survey research on social media and information and communication technology use in Egypt during and since the political crisis surrounding the revolution in Egypt.

Autonomous University of the State of Mexico, Toluca

Organization Type: Academic Institution
Organization Location: Toluca, Mexico

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: collaborated on survey research on social media and information and communication technology use in Mexico during and since the political turmoil surrounding Presidential and Parliamentary elections in Mexico in July 2012

High Institute of Management of Tunis

Organization Type: Academic Institution
Organization Location: Tunis, Tunisia

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: collaborated on survey research on social media and information/communication technology use in Tunisia during and since the political crisis of the revolution.

Internet Archive

Organization Type: Other Nonprofits
Organization Location: San Francisco, CA, USA

Partner's Contribution to the Project:
In-Kind Support

More Detail on Partner and Contribution: The project team is using IA's Archive-It service, specifically the Heritrix crawler and the Wayback machine, for webpage archiving tasks.

University of the Philippines, Diliman

Organization Type: Academic Institution
Organization Location: Philippines

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Collaboration as discussed in the writeup on Dr. Shoemaker's work, including his 3 month visit to the Philippines in the spring.

Have other collaborators or contacts been involved? No

Impacts

What is the impact on the development of the principal discipline(s) of the project?

The IDEAL team built a Hadoop cluster that is used for storing, extracting, analyzing, and visualizing Web and tweet collections. The Hadoop cluster will help speed up processing and analyzing stored collections. This should be an exemplar demonstration of how low-cost equipment can support this type of research, and should guide others with similar interests or with other big data applications.

A novel technique, event focused crawling, for collecting and storing webpages about a specific event was developed. The developed technique will help automate the process of collecting and storing Web collections as well as preparing high quality archives.

We developed an improved technique for location disambiguation in tweets.

We developed an improved technique for topic identification in webpages.

What is the impact on other disciplines?

We have supported related work in library and information science, digital humanities, and other fields interested in working with our collections. Others are likely to use our collections and services as those expand and as we further disseminate results and extend our collaboration.

What is the impact on the development of human resources?

At Virginia Tech, many students learned about and from this project. Students in the spring 2014 offerings of CS4624 (Multimedia, Hypertext, and Information Access) and CS6604 (Digitial Libraries) carried out term projects in groups to help prepare for the fall 2014 offering of Computational Linguistics. Multiple student reports were prepared and made globally accessible through VTechWorks, each listed under Other Publications. Each submission includes midterm and final presentation slides, final report, and all other appropriate deliverables. Students learned not only about related topics in computational linguistics and connected big data issues, but also how to prepare high quality content and upload it to VTechWorks for sharing with others who are interested in learning about these matters.

Also at Virginia Tech, three other students carried out Undergraduate Research (CS2994) in the spring of 2014. One, Joseph Sebastian, learned as he worked to prepare for Computational Linguistics by: 1) helping set up the Hadoop cluster that will be used in the class, and 2) helping load tweet collections into the Hadoop cluster so that content can be used as text for analysis. Two other students (Rohan Kaul and Ishita Ganotra) carried out a different Undergraduate Research project, learning as they helped prepare for Computational Linguistics. Their focus has been on event schema and vocabulary, analyzing the collections to be used in the CL class so templates will be available for each CL group to support text analysis and summarization tasks.

Further, at Virginia Tech, Tarek Kanan is serving this summer as graduate research assistant, supported by some of the wages provided by NSF, to help prepare for the CL course in the fall. His dissertation research relates to work with natural languages, including Arabic and English, as well as computational linguistic techniques and machine learning methods of classification. Thus, he is learning about CL in a way that will aid his further doctoral research. In addition, Xuan Zhang, who played a key role in the spring of 2014 in a related CS6604 student project (see above under Other Publications), will continue to learn about CL since he will be the graduate teaching/research assistant for the new course; this also should inform his doctoral research.

Seungwon Yang defended his Ph.D. dissertation in Dec. 2013 (see Products section).

Sunshin Lee and Mohamed Farag passed their Prelim Exam with work related to the IDEAL project.

What is the impact on physical resources that form infrastructure?

At Virginia Tech, an 11-node Hadoop cluster was set up in the spring of 2014 so it can be used in the CL class in the fall.

A newly upgraded server (with over 25TB of additional storage) will be serving and hosting the IDEAL project.

What is the impact on institutional resources that form infrastructure?

The IDEAL project supported and helped Digital Humanities efforts on campus, especially in the College of Liberal Arts and Human Studies. This led to multiple collaborative projects (building collections about July 4th celebrations since before the Civil War, and about the 1989-1890 Russian Flu, and working with staff at University Libraries on updating the collections about school shootings).

What is the impact on information resources that form infrastructure?

At Virginia Tech, in partial support of the new Computational Linguistics class, there is ongoing work to collect both tweets and webpages to be used as collections that will be analyzed, leading to multiple types of English summaries, to aid the project focus for learning about computational linguistics through projects. There already are many collections, amounting to over 750 million tweets and over 10 TB of webpages.

What is the impact on technology transfer?

Webpage collections are available through the Internet Archive.

Several learning modules were developed that facilitate learning on subjects related to the IDEAL project.

The IDEAL project started collaboration with Altiscale regarding support for distributed processing using Hadoop clusters.

What is the impact on society beyond science and technology?

Public users can access web collections through the IDEAL project website and the IA website. Services are provided that help the public find information about events in our collections. Stakeholder groups also can benefit, that have interest in crises, disasters, tragedies, recovery, and other events (related to communities and governments).

Changes/Problems

Changes in approach and reason for change
Nothing to report.

Actual or Anticipated problems or delays and actions or plans to resolve them
Nothing to report.

Changes that have a significant impact on expenditures
Nothing to report.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals
Nothing to report.

Significant changes in use or care of biohazards
Nothing to report.