Reducing Information Overload in Emergencies by Detecting Themes in Web Content

Jorge H. Roman, Linn Marks Collins, Ketan K. Mane, Mark L. B. Martinez, Carolyn E. Dunford, James E. Powell, Jr.
Los Alamos National Laboratory
{jhr, linn, kmane, mlbm, jepowell}@lanl.gov

ABSTRACT

Information on the Web has become increasingly important in disaster response. Yet much of this information is redundant. We are creating a suite of electronic knowledge management (eKM) tools that can be used to reduce by an order of magnitude the information that people need to read in order to gain and maintain awareness of web content during emergencies. In this paper, we describe research-in-progress on developing these tools and applying them to web content from organizations' websites and individuals' blogs. Results so far indicate that organizations' websites and individuals' blogs provide redundant coverage of general issues and that each provides additional information about specific issues. By using the tools we are developing, responders and victims will be able to quickly gather an overview of general issues derived from many websites, then learn more about specific issues by navigating to a few websites.

Keywords

Emergency information system, information overload, theme detection, information extraction, web crawl, unstructured text

INTRODUCTION

The amount of information on the Web about emergencies exceeds any individual's or group's ability to read and absorb in a timely manner. For example, a search on Google for information about "California wildfires" 2007 on November 20, 2007 yielded 1,640,000 results.

We are creating a suite of electronic knowledge management (eKM) tools for automatically collecting web content and identifying key themes in order to reduce this information to a manageable set of concepts. Our goal is to reduce by an order of magnitude the information that users need to read in order to gain and maintain awareness of web content during emergencies.

Unstructured text processing typically focuses on search algorithms. Tools analyze sequences of strings (search patterns) and the search results consist of links to documents containing these strings, sometimes augmented by ranking strategies, e.g. Google's page rank algorithm (Kleinberg, 1999).

Meta search engines provide value by searching multiple search engines, and reprocessing the list of ranked documents and some snippet of the original text (Osiński and Weiss, 2004; Stefanowski and Weiss, 2003). Additional value is provided by the Carrot2 Meta Search engine which clusters results. The Carrot2 algorithm is based on the Extensible Federated Search Server (http://www.etools.ch/searchInfo.do).

These approaches are optimized for finding every document with a given string sequence rather than just those documents where the string sequence is a key concept. Furthermore, these approaches are optimized for searching for single concepts, as opposed to sets of concepts.

If we are interested in more complex searches, for trends or concept maps, then other tools need to be used. Autonomy is a leading provider of commercial knowledge management tools that can, among other things, dynamically build taxonomies (Marks, 2007; http://www.autonomy.com/content/Technology/index.en.html). IBM takes a different approach by using an open standard framework where Unstructured Information Modules can be plugged in to work together in solving knowledge management needs (Ferrucci and Lally, 2004;

http://domino.research.ibm.com/comm/research_projects.nsf/pages /uima.knowledgeRush.html). Other technologies use the Web 2.0 metadata approach (Hendler and Golbeck, 2008). One drawback of these approaches is the need for a large set of training material and/or tailored handcrafting of the modules to address particular content.

In contrast, the approach we take is that each document contains knowledge that can be leveraged. Each document is analyzed and a hierarchical set of knowledge signatures or concepts is created (kSigs). A set of kSigs can be merged to construct a taxonomy. Taxonomy operations can reveal high-level patterns on large sets of unstructured text. If the sets are constructed for some time interval, then we can compute emerging trends by comparing the generated taxonomies from the previous time interval.

Taxonomy operations typically involve hand comparisons. Our approach uses automated operations by integrating commercial tools, open-source tools, custom-made modules, and visualization techniques to reduce 10,000+ documents to a set of key concepts.

Once web content has been collected and prepared (e.g. converted from PDF to HTML), eKM consists of five operations. (Figure 1)

1. kSig computation: The first operation involves automatically summarizing conceptual content. The body of a text document is reduced to a hierarchy of concepts called a knowledge Signature (kSig). This operation results in a 10:1 reduction of content (original text : kSig). It also generates navigational aids from the kSig to the original content.
2. Collection management: The second operation involves clustering into profiles. The profiles can be manually generated or derived from the content.

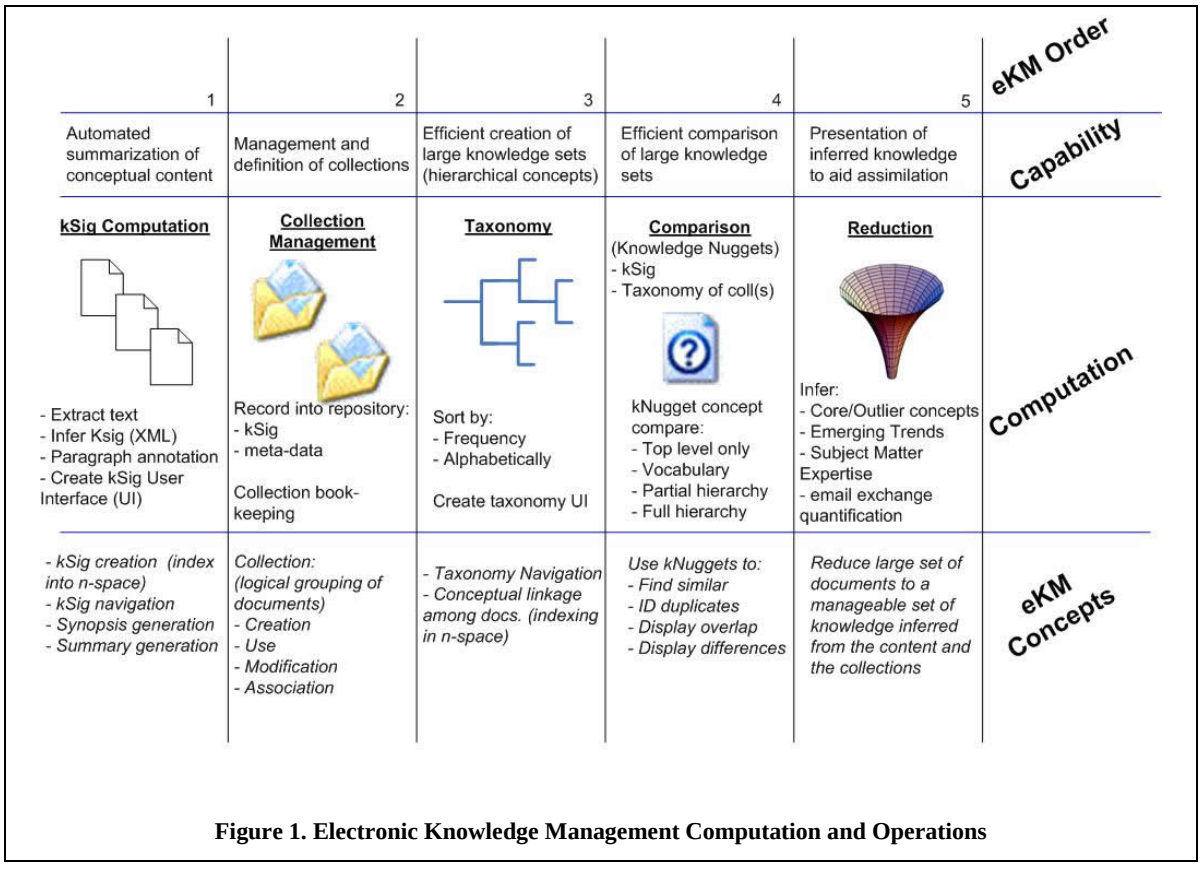


Figure 1. Electronic Knowledge Management Computation and Operations

3. Taxonomy creation: Once a collection of related content is identified, a taxonomy can be computed by creating a union of the kSigs. Taxonomy creation generates an index and navigational aids to the content.

4. Comparison: Once taxonomies have been created for a set of documents, they can be compared to identify similar and dissimilar web content and trends over time.
5. Reduction: After kSigs have been computed for content and taxonomies have been sorted by frequency of terms, we can make inferences based on the frequency of terms: for example, that one aspect of emergency response is not addressed on some websites, but is addressed on others.

METHOD

In this paper, we report on results gleaned from testing several of these tools on web content from organizations' websites and individuals' blogs during the California wildfires in the fall of 2007.

Content collection

To collect content for analysis, we used one custom-made web crawler and JoBo, an open-source web crawler which downloads complete websites (http://www.matuschek.net/jobo/). Using the custom-made web crawler, we crawled the Federal Emergency Management Agency (FEMA) web pages about California wildfires (http://www.fema.gov/news/eventnews.fema?id=9045). Using Jobo, we crawled the State of California Department of Forestry and Fire Protection (DFFP) website (http://www.fire.ca.gov) and the bloggersblog.com website (http://www.bloggersblog.com/2007californiawildfires/).

Content preparation

To prepare content that has been downloaded we used Xpdf, an open-source tool for PDF file manipulation which extracts text and images from PDF files in order to generate HTML files (http://www.foolabs.com/xpdf/).

kSig computation

To compute kSigs for individual pages from the downloaded content (the DFFP website, the FEMA website, and bloggersblog.com), we used the theme extractor from a commercial product, CIRI Lab's Knowledge Object Suite (KOS) (http://www.cirilab.com). Through an algorithm, the theme extractor takes sentences and identifies key themes. We configured this tool to generate a kSig of shape 6, 2, 2 for each file, where the three numbers indicate the number of concepts at each of three levels: 6 concepts at the top level, 2 concepts at the second level, and 2 concepts at the third level. Then these themes are weighted and combined to compute a kSig for the entire page.

Collection management

To manage the collection, we stored the kSigs in Apache Derby, an open-source embedded relational database engine (http://db.apache.org/derby/). To automatically create a database schema and upload the data to the Derby database, we used custom-made tools. These tools examine the kSig directory structure for a given website and generate files with data pointers and instructions for creating database tables and uploading the data.

Taxonomy creation

To create a taxonomy of the content, we indexed the kSigs in the database using Apache Lucene, an open-source search engine (http://lucene.apache.org/java/docs/), and created a union of the kSigs. We used a custom-made taxonomy viewer to view the results.

Comparison

To compare themes, we used custom-made tools for analyzing unions and intersections of kSigs, specifically:

- The intersection of kSigs for the DFFP and FEMA websites, which yields a list of themes that are common to both organizations' websites
- The union of kSigs for the bloggersblog.com website (which is, itself, a compilation of blogs), which yields a list of themes that are common to individuals' blogs
- The intersection of kSigs for the organizations' websites and the individuals' blogs, which yields a list of themes that are common to both kinds of web content

- The intersection of kSigs for each of the websites – the DFFP, FEMA, and bloggersblog.com websites – which yields a list of themes that are common to all three websites

RESULTS

For organizations' websites, we downloaded 1681 files from the State of California Department of Forestry and Fire Protection (DFFP) website. These files yielded 10473 kSigs or unique themes. We downloaded 53 files from the Federal Emergency Management Agency (FEMA) website. These files yielded 887 kSigs or unique themes.

For individuals' blogs, we downloaded one file from the bloggersblog.com website, consisting of several individuals' blogs. This file yielded 42 kSigs or unique themes. (Figure 2) Since bloggersblog.com is, itself, a compilation of blogs, these 42 kSigs can be considered the themes that are common to individuals' blogs.

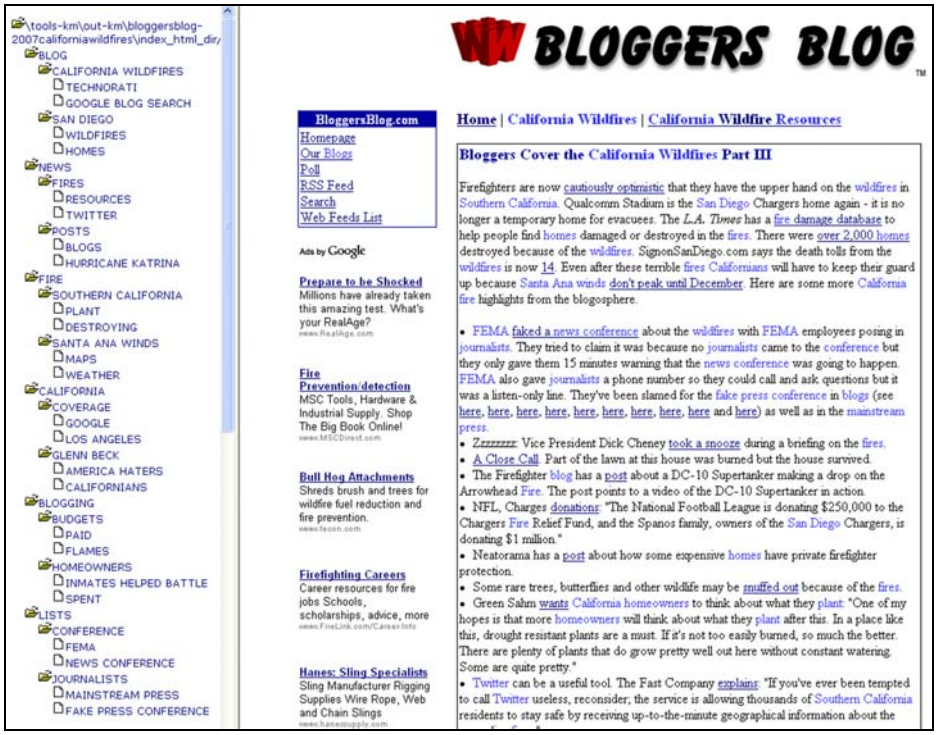


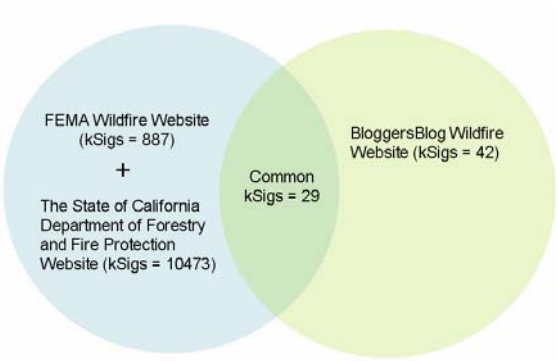
Figure 2. Knowledge signatures (left) for bloggersblog.com web content (right)

It took approximately three hours to collect content from the three websites. It took approximately two-and-a half hours to compute kSigs for the content using a Windows XP, 1 CPU, 2GB machine.

The intersection of kSigs for the DFFP and FEMA websites yielded a list of 310 themes that are common to both organizations' websites.

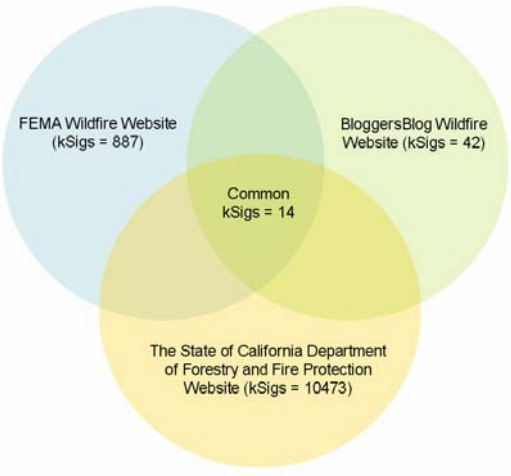
The intersection of 310 themes for the organizations' websites and 42 for the individuals' blogs yielded a list of 29 themes that are common to organizations' websites and individuals' blogs. (Figure 3)

The intersection of kSigs for the DFFP, FEMA, and bloggersblog.com websites yielded a list of 14 themes that are common to all three websites. (Figure 4)



Blog	Fires	Paid
Budgets	Flames	Plant
California	Homeowners	Resources
California wildfires	Homes	San Diego
Californians	Hurricane Katrina	Santa Ana Winds
Conference	Lists	Southern California
Coverage	Los Angeles	Spent
Destroying	Maps	Weather
FEMA	News	Wildfires
Fire	News conference	

Figure 3. Themes common to organizations' websites and individuals' blogs (29)



California	Homeowners	San Diego
Californians	Homes	Southern California
Coverage	Los Angeles	Spent
Fire	Paid	Wildfires
Fires	Resources	

Figure 4. Themes common to all three websites (14)

DISCUSSION

Using our suite of tools, we have been able to:

- Extract 11402 kSigs from a set of 1735 files downloaded from the web
- Determine two sets of common themes from the set of 11402 kSigs, consisting of 14 and 29 terms

These sets of kSigs provide an overview of web content. By viewing them, users can quickly gain an overview of the themes that are most important in unstructured text downloaded from the web. By using the navigational aids, they can visit the source web pages and read more.

In order to analyze the content further, users can analyze the union and intersection of kSigs. For example, the theme of money (kSigs = budgets, paid, spent) is common to both organizations' and individuals' web content. (The kSig "coverage" might refer to either insurance coverage or news coverage.) Of the 11402 kSigs, only 29 are common to both kinds of content and at least three – possibly four – of these relate to money. This suggests that issues regarding money were extremely important in this emergency.

Conversely, by analyzing the kSigs or themes that are not common to the organizations' websites and the individuals' blogs, users can gain a different, finer-grained perspective of issues and concerns during emergencies. The themes of help provided by prison inmates (kSig = inmates helped battle) and social networking tools (kSig = Twitter) occur only in individuals' blogs, for example.

Thus, users can differentiate between themes that are important to victims and responders, as a whole, vs. those that are important to a subset of the population. As new web content becomes available, they can determine how themes change over the course of an emergency or its aftermath.

The demonstrated capability is currently limited to processing thousands of files in a few hours with minimal human intervention. To be able to handle web content during an emergency, ideally hundreds of thousands of files would be processed automatically near real time. It is likely that a taxonomy for a large collection of documents would contain some 30,000 concepts. This, in turn, would require additional functionality, such as concept clustering algorithms, in order to reduce the concepts to a more manageable set.

CONCLUSION

Using theme extraction and other electronic Knowledge Management (eKM) tools and techniques, themes can be automatically extracted from unstructured text and organized into a set of hierarchical themes that represent the key concepts in the text. Sets of themes can be organized into a taxonomy representing thousands of documents. With performance improvements, these tools will be able to handle more data, more quickly.

Users can quickly grasp themes and access the original information as needed. While these kinds of analyses and navigational aids may not help firefighters in the field, they may be important to public information officers as they attempt to disseminate information and to government agencies as they try to determine priorities.

In certain kinds of emergencies, such as public health emergencies where scientific and medical information needs to be integrated with local reports of symptoms, eKM tools and techniques can help experts and the public gain an overview of a complex situation quickly and then refer to source material for scientific and medical details (Collins, Powell, Dunford, Mane, and Martinez, 2008). As the emergency evolves over time, users can track changes in themes.

Themes and taxonomies extracted from web content during one emergency may help in preparation and training for future emergencies. In addition, they may help in improving emergency information systems, tools, and technologies by making it possible to focus web crawlers and to improve taxonomies for organizing emergency information.

REFERENCES

1. Apache Derby project: http://db.apache.org/derby/
2. Apache Lucene: http://lucene.apache.org/java/docs/
3. Autonomy: http://www.autonomy.com/content/Technology/index.en.html
4. bloggersblog.com: http://www.bloggersblog.com/2007californiawildfires/

5. CIRI Lab Instigator: http://www.cirilab.com
6. Collins, L.M., Powell, J.E., Jr., Dunford, C.E., Mane, K.M., and Marrtinez, M.L.B. (2008) Emergency Information Synthesis and Awareness Using E-SOS, *Proceedings of the 5th International ISCRAM Conference*, Washington, DC, USA.
7. eTools Metasearch Engine: http://www.etools.ch/searchInfo.do
8. Federal Emergency Management Agency (FEMA): http://www.fema.gov/news/eventnews.fema?id=9045
9. Ferrucci, D. and Lally, A. (2004) UIMA: an architectural approach to unstructured information processing in the corporate research environment, *Natural Language Engineering*, 10, 3-4, 327-348.
10. Hendler, J. and Golbeck, J. (2008) Metcalfe's law, Web 2.0, and the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web 2008*6, 1, 14-20.
11. JoBo: http://www.matuschek.net/jobo/
12. Kleinberg, J.M. (1999) Authoritative Sources in a Hyperlinked Environment, *Journal of the ACM*, 46, 5, 604-632.
13. Marks, P. (2007) Improving the Search for Intelligence, *New Scientist Magazine*, February 10, 2007, 2590.
14. Office of the Governor of the State of California: http://www.calfires.com
15. Osinski, S. and Weiss, D. (2004) Carrot2: An Open Source Framework for Search Results Clustering, Poster, *26th European Conference on Information Retrieval*, Sunderland, UK.
16. State of California Department of Forestry and Fire Protection (DFFP): http://www.fire.ca.gov
17. Stefanowski, J. and Weiss, D. (2003) Carrot2 and Language Properties in Web Search Results Clustering, *Lecture Notes in Artificial Intelligence: Advances in Web Intelligence, Proceedings of the First International Atlantic Web Intelligence Conference*, Madrid, Spain, 2663, 240-249.
18. Technorati: http://charts.technorati.com/chart/california+wildfires?language=en&authority=a4
19. Unstructured Information Management Architecture (UIMA): http://domino.research.ibm.com/comm/research_projects.nsf/pages/uima.knowledgeRush.html
20. Xpdf: http://www.foolabs.com/xpdf/