VAST to Knowledge: Combining tools for exploration and mining

Loretta Auvil[1]*, Xavier Llorà[2]†, Duane Searsmith[1]‡, and Kelly Searsmith[1]§

[1]Automated Learning Group
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

[2]Data Intensive Technologies and Applications
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

ABSTRACT

The investigation of the VAST Contest collection provided a valuable test for text mining techniques. Our group has focused on creating analytical tools to unveil relevant patterns and to aid with the content navigation in such text collections. Our results show how such an approach, in combination with visualization techniques, can ease the discovery process especially when multiple tools founded on the same approach to data mining are used in complement to and in concert with one another.

Keywords: Text mining, digital libraries, information visualization, visual analytics, knowledge discovery.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces; I.2.6 [Artificial Intelligence]: Learning I.2.7 [Artificial Intelligence]: Natural Language Processing

1 INTRODUCTION

Text mining and analytical capabilities have been specially designed to operate in and capitalize upon the complexity of rich natural language domains of very large stores of text and multimedia documents. The VAST collection therefore represents an excellent case study for our tools. Our efforts focused on exploring relevant patterns hidden in the VAST collection; using them to identify key players, frequent repetitions, key concepts, and their relationships; and automatically ranking relevant excerpts. Each of these techniques was directed toward helping analysts to navigate the collection in an aided and meaningful way.

2 BACKGROUND

Our investigation employed four different tools: D2K (Data To Knowledge), FeatureLens, RiverGlass ReCon™, and DISCUS. Each tool was selected for its different analytical functions, but together these tools share the same basic approach: finding key terms and key links that bridge high frequency clusters, and pointing to interesting transitions between the concepts described by those clusters.

D2K: Used here in several ways, D2K (http://alg.ncsa.uiuc.edu) was developed by the Automated Learning Group at NCSA to serve as a rapid, flexible data mining and machine learning system. D2K integrates analytical data mining methods for prediction, discovery, and deviation detection with data and information visualization tools. D2K offers a visual programming environment that allows users to connect programming modules together to build data mining applications; it also supplies a core set of modules, application templates, and a standard API for software component development.

FeatureLens: FeatureLens[1] [1, 2] provides an interface for exploring and visualizing features in collections of text documents. It allows researchers to explore frequent patterns, from frequently used words to frequent patterns of ngrams. FeatureLens integrates the results of text-mining algorithms into a meaningful representation of a collection. Features can be compared, and occurrences of the patterns are shown in the text. To help users in finding interesting patterns, FeatureLens highlights features that have specific patterns of use in the collection (e.g., increasing, decreasing, spike behavior). FeatureLens was created in Spring 2007.

RiverGlass ReCon™: RiverGlass[2] ReCon™ is used to find, collect, and analyze text information from the web and internal document repositories using text analytics. The tool employs a variety of techniques to perform this analysis: semantic technology, domain ontologies, natural language processing, document summarization, information extraction, text classification, and clustering for relevance feedback.

DISCUS: DISCUS[3] encompasses several analytics tools. Summarizer is used to rank the sentences and words of a collection and collection subsets [3]. The ranking is based on a mutually reinforcing relationship between sentences and terms: important sentences include many important terms, and conversely, important terms are included by many important sentences. Concept Map uses a chance discovery technique called KeyGraph, which provides a visual map of the contents of the collection. This visualization has been widely used as a tool for human innovation and creativity in on-line scenarios for market trend detection [4].

3 ANALYTICAL PROCESS

Our process started by using D2K to perform frequent pattern analysis. The VAST document collection contained duplicates, causing long patterns. Tight document clustering with D2K revealed this, and enabled the removal of duplicate documents from the collection. Figure 1 shows the D2K environment and the itinerary (workflow) used to perform the clustering. Once the collection was cleaned, an examination of pictures in the document collection led us to a search for content related to *"chinchilla(s)"* and *"chin(s),"* because this animal was a recurrent topic of photos and blogs included in the document collection.

RiverGlass's ReCon™ clustered the retrieved documents obtained after the previous search and allowed us to interactively read them based on their similarity. We could actively repeat the search-and-cluster approach to form a subset, and then cluster the subset while refining the view of documents related to a particular topic search. Figure 2 shows a display of the clustering. ReCon™ was also used to perform entity extraction, automatically revealing entities that co-occured in the same documents, and thus identifying people and organizations that may have relationships. ReCon™

*e-mail: lauvil@uiuc.edu
†e-mail: xllora@ncsa.uiuc.edu
‡e-mail: dsears@ncsa.uiuc.edu
§e-mail: ksearsmi@ncsa.uiuc.edu

[1] http://www.cs.umd.edu/hcil/textvis/featurelens
[2] http://www.riverglassinc.com
[3] http://www-discus.ge.uiuc.edu

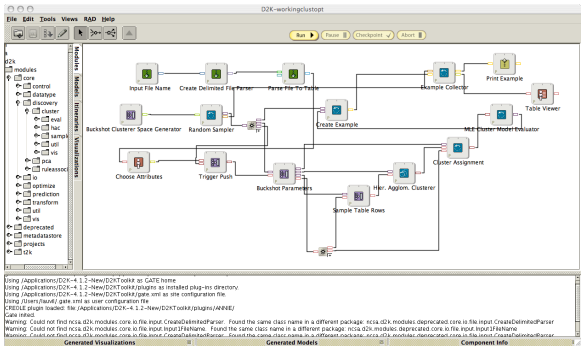

Figure 1: D2K showing itinerary for clustering of the vast document collection.

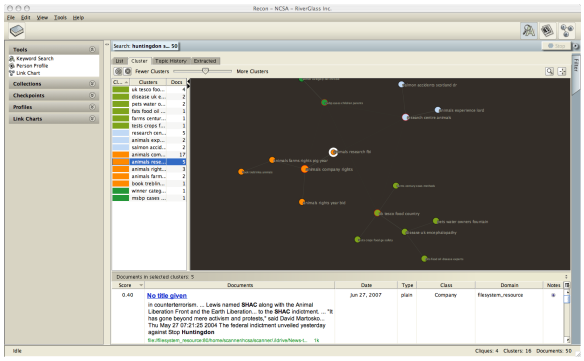

Figure 2: RiverGlass ReCon™ clustering higlights node "animal research fbi" and its list of documents shown below.

uses ontologies built to target terrorism, violent crimes, and narcotics. ReCon™'s results helped us to lay out the map of players involved in the VAST collection.

With such information at hand, we used FeatureLens to perform word and 3 gram pattern analysis inside the collection database created by D2K. FeatureLens enables users to find meaningful co-occurrences of text patterns and their evolution by visualizing them within and across documents in the collection. Features can be compared, and occurrences of the patterns are shown in the text. Each pattern is assigned a different color, and when a document contains one of the selected patterns, the color saturation of the line reflects the score of the pattern in the documents. FeatureLens's value was its ability to identify patterns of terms for further analysis. For instance, a reference to "bombings" led to the discovery of bombing sites: "Chiron" and "Shaklee." Figure 3 shows analysis using FeatureLens.

To help summarize key concept relations and document facts, we used DISCUS. The sentence ranking mechanism employed by DISCUS extracted relevant portions of text, which, ultimately, provided a useful global summarization of the collection's main narratives. This summarization helped articulate the findings obtained using ReCon™ and FeatureLens. Moreover, DISCUSs innovative visualization techniques helped lay out concept maps. Such a visualization works by computing high-frequency terms and the more frequent links among them–links are computed inside sentences. Figure 4 shows a concept map, created using a document subset provided by the tools mentioned above, highlighting the keywords "huntingdon life science," "life science," and "animal cruelty."

4 Conclusion

Using four different, but complementary tools (D2K, FeatureLens, RiverGlass ReCon™, and DISCUS) we have been able to mine and explore the VAST contest collection. We unveiled relevant patterns,

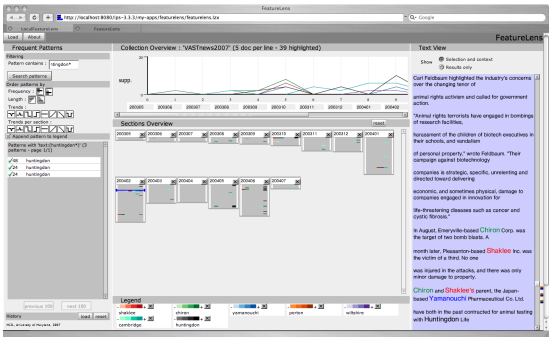

Figure 3: FeatureLens identifying co-occurrence of words "Shaklee," "Chiron," "Yamanouchi," "Porton," "Wiltshire," "Cambridge" and "Huntingdon." The collection is divided by month. Trend graphs are shown above and relevant highlighted documents on the right.

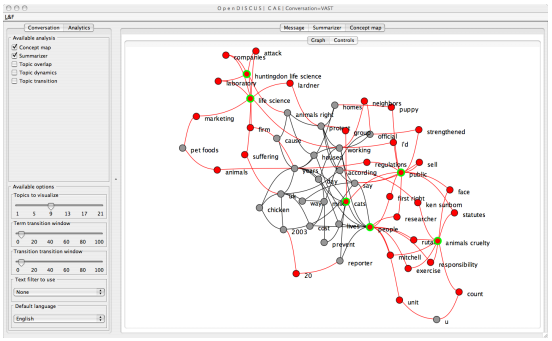

Figure 4: DISCUS Concept Map showing how low frequency words and links bridge high frequency clusters, pointing to interesting transitions between concepts.

which met our content navigation needs. Facts were gathered, concept maps drawn, and narrative articulation provided, proving the viability of combining these four applications to address the category of problem described by the VAST contest.

Acknowledgements

This work was sponsored by the AFOSR (F49620-03-1-0129), and NSF (IIS-02-09199). Thanks to Anthony Don and other developers of FeatureLens, RiverGlass Inc. for providing access to their tool, and Noriko Imafuji Yasui for help with DISCUS.

References

[1] T. Clement, A. Don, C. Plaisant, L. Auvil, G. Pape, and V. Goren. Something that is interesting is interesting then: Using text minig and visualizations to aid interpreting repetition in Gertrude Stein's The Making of Americans. In *Proceedings of the Digital Humanities Conference*, pages 40–44, 2007.

[2] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant. Discovering interesting usage patterns in text collections: Integrating text mining with visualization. In *Proceedings 2007 ACM Conference on Information and Knowledge Management (to appear)*, 2007.

[3] X. Llorà, D. E. Goldberg, Y. Ohsawa, N. Matsumura, Y. Washida, H. Tamura, M. Yoshikawa, M. Welge, L. Auvil, D. Searshmith, K. Ohnishi, and C. J. Chao. Innovation and creativity support via chance discovery, genetic algorithms, and data mining. *New Mathematics and Natural Computation*, 2(1):85–100, 2006.

[4] N. I. Yasui, X. Llorà, D. E. Goldberg, Y. Washida, and H. Tamura. Delineating topic and discussant transitions in online collaborative environments. In *Proceedings of the International Conference on Enterprise Information Systems*, volume AIDSS, pages 14–21, 2007.