FutureLens

Gregory Shutt
Andrey Puretskiy
Michael W. Berry

Department of Electrical Engineering and Computer Science
The University Of Tennessee

November 20, 2008
Knoxville, TN

Motivation

Visualization can be a very powerful tool in data mining when applied properly. It can lead to very quick knowledge discovery as it allows for a big picture overview of overwhelmingly large amounts of data. In this particular task, a tool was to be designed that would take a large set of SGML documents and present them to the end user in such a manner to facilitate quickly finding interesting patterns and aid in knowledge discovery. The additional requirement that output from a non-negative tensor factorization tool be directly input to the program was applied. The non-negative tensor factorization produced twenty-five different groups of fifteen relevant specific entities and thirty-five relevant terms. An entity in this case is a specific type of term such as a person, location, or organization [4] [5].

Background

Many of the concepts and ideas of this project stem from FeatureLens, a University of Maryland text and pattern visualization program [1]. FeatureLens allows the user to explore frequently occurring terms or patterns in a collection of documents. Connections between these frequent terms and the dates at which they appear in the set of documents can quickly be visualized and investigated. A screen shot of FeatureLens is shown below in Figure 1.

Figure 1: FeatureLens

While FeatureLens may sound suitable for the given task, it is not without its shortcomings. For one, its design is rather complex as it requires a MySQL database server, an HTTP

server, and an Adobe Flash enabled web browser to function properly. As such, it is not a trivial task to set up an instance of FeatureLens from scratch and may take an inexperienced user a significant amount of time to get started. Data sets must be be parsed and stored in the database, an operation that an end user cannot perform so examining arbitrary data sets is out of the question. In implementing the architecture of FeatureLens, the designers chose to use a variety of languages–Ruby for the back end, XML to communicate between the front end and back end, and OpenLaszlo for the interface. Because of this variety in languages adapting and modifying FeatureLens would prove quite difficult. Responsiveness of the interface also tends to degrade to the point that it impacts usability when given even the simplest of tasks. Clearly a better solution was needed.

Features

FutureLens is a text visualization tool that implements much of the functionality of FeatureLens while adding a few necessary missing features. It is written in the Java programming language using the Standard Widget Toolkit so it is not only cross platform but uses native widgets where possible to maintain a consistent look and feel with the platform it is being run on. For end users not familiar with the program, FutureLens has a built in demo feature that demonstrates its basic functionality. An example of FutureLens running on Mac OS X is shown below in Figure 2.

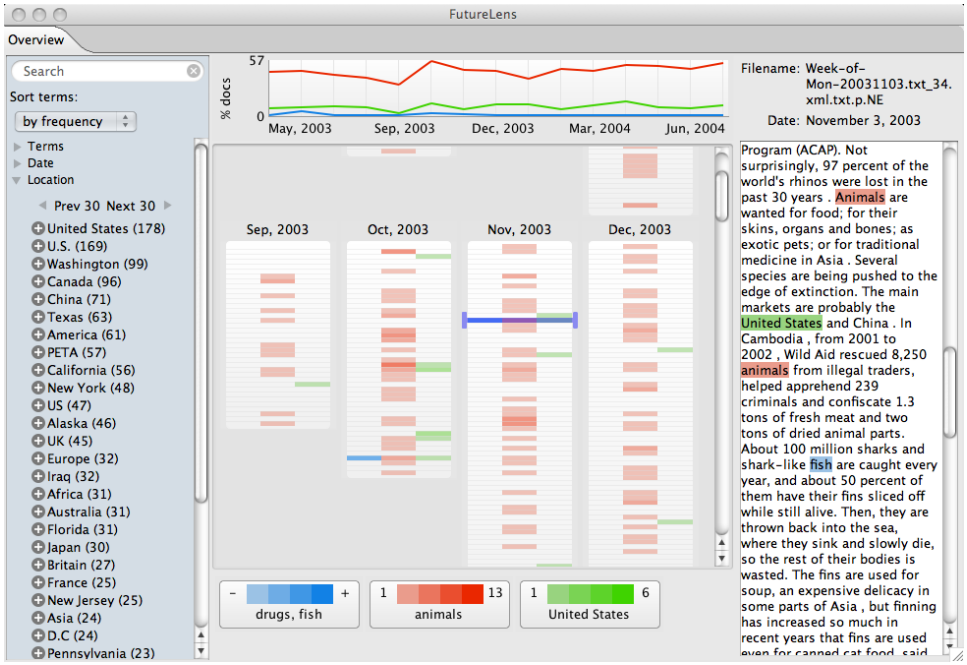

Figure 2: FutureLens

All the basic functionality of FutureLens can be seen in this example. The boxes along the bottom show the terms that are currently being investigated. The intensity of the color hints at the concentration of the term throughout the documents. A graph of the percentage of documents containing the term versus time is shown at the top, while the raw text of the selected document is shown to the right with the selected terms highlighted in the appropriate color. Multiple terms can be combined into extended patterns easily by dragging and dropping. While this presents an excellent overview of the data, it is possible to load the output of a data mining tool. An example of this is shown below in Figure 3.

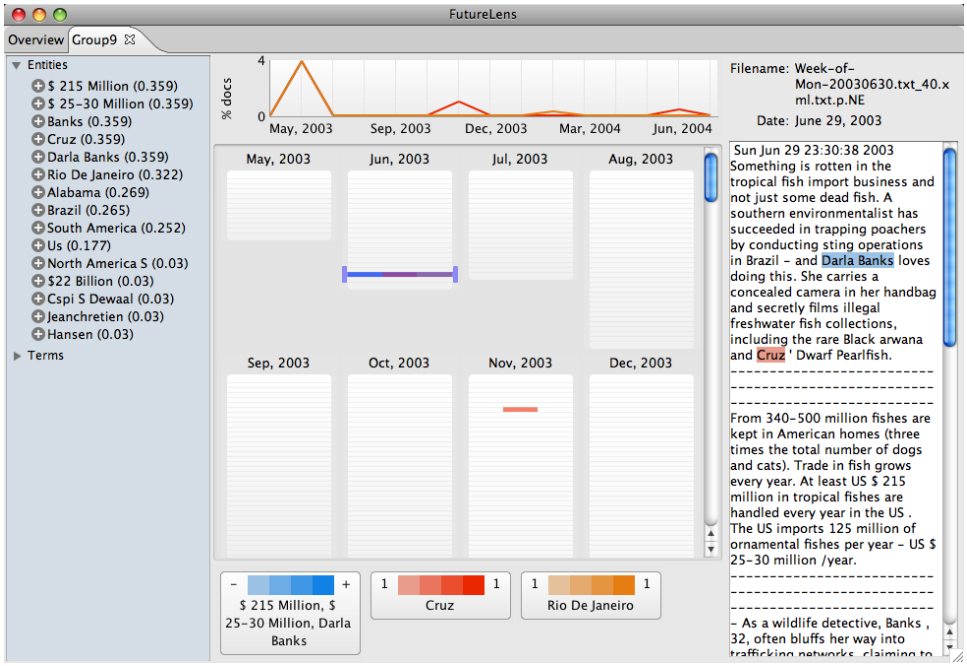

Figure 3: FutureLens With Group File

Here a file containing pertinent terms output from a tensor factorization tool has been loaded as a separate view into FutureLens. The view is nearly identical to the overview. However, the list of terms has been limited to only what was contained in the input file. This allows the user to quickly view the different clusters of entities through time [5].

Future Work

While FutureLens implements all the required features for the given task, it can still stand some improvement. It works well for evidence generation but it has no automation for any type of scenario discovery. Methods that locate interesting features in the data set could be added to create a single analysis tool. As it stands now, the output of data mining methods

must manually be entered into the program. Eliminating this human interaction would greatly increase the efficiency of scenario discovery. The program can also be extended to support a dynamic data set if such a need might arise.

References

[1] Exploring and visualizing frequent patterns in text collections with FeatureLens. `http://www.cs.umd.edu/hcil/textvis/featurelens`. Visited November 2008.

[2] The MONK Project Wiki. `https://apps.lis.uiuc.edu/wiki/display/MONK/The+MONK+Project+Wiki`. Last edited August 2008.

[3] Brett W. Bader, Michael W. Berry, and Murray Brown. Discussion tracking in Enron email using PARAFAC. In M.W. Berry and M. Castellanos, editors, *Survey of Text Mining II: Clustering, Classification, and Retrieval*, pages 147–163. Springer-Verlag, London, 2008.

[4] Brett W. Bader, Michael W. Berry, and Amy N. Langville. Nonnegative matrix and tensor factorization for discussion tracking. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory, Applications, and Visualization*. Chapman & Hall / CRC Press, 2008.

[5] Brett W. Bader, Andrey A. Puretskiy, and Michael W. Berry. Scenario discovery using nonnegative tensor factorization. In Jose Ruiz-Shulcloper and Walter G. Kropatsch, editors, *Progress in Pattern Recognition, Image Analysis and Applications, Proceedings of the Thirteenth Iberoamerican Congress on Pattern Recognition, CIARP 2008, Havana, Cuba, Lecture Notes in Computer Science (LNCS) 5197*, pages 791–805. Springer-Verlag, Berlin, 2008.

[6] A. Don, E. Zhelev, 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. *HCIL Technical report 2007-08*, May 2007.