# InfoMagnets: Making Sense of Corpus Data

**Jaime Arguello**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15216
jarguell@andrew.cmu.edu

**Carolyn Rosé**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15216
cprose@cs.cmu.edu

## Abstract

We introduce a new interactive corpus exploration tool called InfoMagnets. InfoMagnets aims at making exploratory corpus analysis accessible to researchers who are not experts in text mining. As evidence of its usefulness and usability, it has been used successfully in a research context to uncover relationships between language and behavioral patterns in two distinct domains: tutorial dialogue (Kumar et al., submitted) and on-line communities (Arguello et al., 2006). As an educational tool, it has been used as part of a unit on protocol analysis in an Educational Research Methods course.

## 1 Introduction

Exploring large text corpora can be a daunting prospect. This is especially the case for behavioral researchers who have a vested interest in the latent patterns present in text, but are less interested in computational models of text-representation (e.g. *the vector-space model*) or unsupervised pattern-learning *(*e.g. *clustering)*. Our goal is to provide this technology to the broader community of learning scientists and other behavioral researchers who collect and code corpus data as an important part of their research. To date none of the tools that are commonly used in the behavioral research community, such as HyperResearch, MacShapa, or Nvivo, which are used to support their corpus analysis efforts, make use of technology more advanced than simplistic word counting approaches. With InfoMagnets, we are working towards bridging the gap between the text-mining community and the corpus-based behavioral research community. The purpose of our demonstration is to make the language technologies community more aware of opportunities for applications of language technologies to support corpus oriented behavioral research.

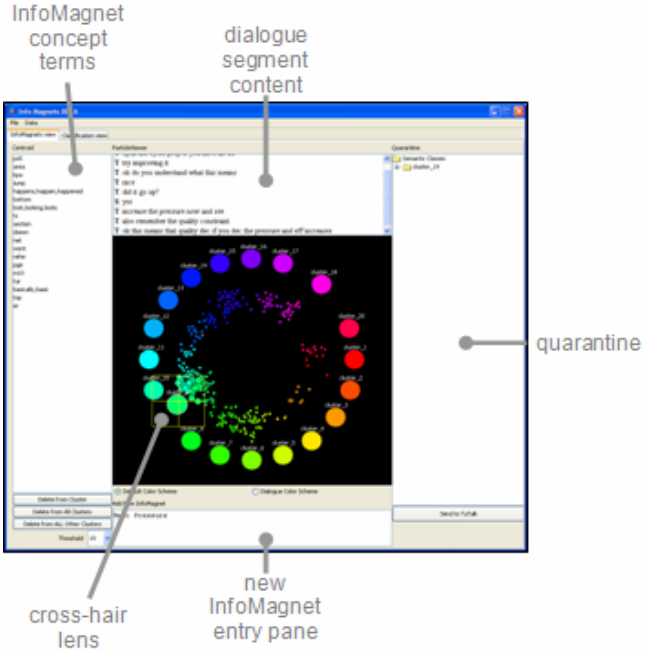

Figure 1: InfoMagnets Screenshot

InfoMagnet's novelty is two-fold: First, it provides an intuitive visual metaphor that allows the user to get a sense of their data and organize it for easy retrieval later. This is important during the sense making stage of corpus analysis work just before formal coding scheme development begins. Secondly, it allows the user to interact with clustering technology, and thus influence its behavior, in effect introducing human knowledge into the clustering process. Because of this give and take between the clustering technology and the human

influence, the tool is able to achieve an organization of textual units that is not just optimal from an algorithmic stand-point, but also optimal for the user's unique purpose, which non-interactive clustering algorithms are not in general capable of achieving.

Using visual metaphors to convey to the user proximity and relations between documents and automatically generated clusters is not a new technique (Chalmers and Chitson, 1992; Dubin, 1995; Wise et al., 1995; Leuski and Allan, 2000; Rasmussen and Karypis, 2004). InfoMagnet's novelty comes from giving the user more control over the ultimate clustering organization. The user is able to incrementally influence the formation and reorganization of cluster centroids and *immediately* see the effect on the text-to-cluster assignment. Thus, the user can explore the corpus in more effective and meaningful ways.

In what follows, we more concretely elaborate on InfoMagnet's functionality and technical details. We then motivate its usability and usefulness with a real case study.

## 2 Functionality

Exploring a textual corpus in search of interesting topical patterns that correlate with externally observable variables is a non-trivial task. Take as an example the task of characterizing the process by which students and tutors negotiate with one another over a chat interface as they navigate instructional materials together in an on-line exploratory learning environment. A sensible approach is to segment all dialogue transcripts into topic-oriented segments and then group the segments by topic similarity. If done manually, this is a challenging task in two respects. First, to segment each dialogue the analyst must rely on their knowledge of the domain to locate where the focus of the dialogue shifts from one topic to the next. This, of course, requires the analyst to know what to look for and to remain consistent throughout the whole set of dialogues. More importantly, it introduces into the topic analysis a primacy bias. The analyst may miss important dialogue digressions simply because they are not expected based on observations from the first few dialogues viewed in detail. InfoMagnets addresses these issues by offering users a constant bird's eye view of their data. See Figure 1.

As input, InfoMagnets accepts a corpus of textual documents. As an option to the user, the documents can be automatically fragmented into topically-coherent segments (referred to also as *documents* from here on), which then become the atomic textual unit[1]. The documents (or topic segments) are automatically clustered into an initial organization that the user then incrementally adjusts through the interface. Figure 1 shows the initial document-to-topic assignment that InfoMagnets produces as a starting point for the user. The large circles represent InfoMagnets, or topic oriented cluster centroids, and the smaller circles represent documents. An InfoMagnet can be thought of as a set of words representative of a topic concept. The similarity between the vector representation of the words in a document and that of the words in an InfoMagnet translate into attraction in the two-dimensional InfoMagnet space. This semantic similarity is computed using Latent Semantic Analysis (LSA) (Landauer et al., 1998). Thus, a document appears closest to the InfoMagnet that best represents its topic.

A document that appears equidistant to two InfoMagnets shares its content equally between the two represented topics. Topics with lots of documents nearby are popular topics. InfoMagnets with only a few documents nearby represent infrequent topics. Should the user decide to remove an InfoMagnet, any document with some level of attraction to that InfoMagnet will animate and reposition itself based on the topics still represented by the remaining InfoMagnets. At all times, the InfoMagnets interface offers the analyst a bird's eye view of the entire corpus as it is being analyzed and organized.

Given the automatically-generated initial topic representation, the user typically starts by browsing the different InfoMagnets and documents. Using a magnifying cross-hair lens, the user can view the contents of a document on the top pane. As noted above, each InfoMagnet represents a topic concept through a collection of words (from the corpus) that convey that concept. Selecting the InfoMagnet displays this list of words on the left pane. The list is shown in descending order of importance with respect to that topic. By browsing each InfoMagnet's list of words and browsing

[1] Due to lack of space, we do not focus on our topic-segmentation algorithm. We intend to discuss this in the demo.

nearby documents, the user can start recognizing topics represented in the InfoMagnet space and can start labeling those InfoMagnets.

InfoMagnets with only a few neighboring documents can be removed. Likewise, InfoMagnets attracting too many topically-unrelated documents can be split into multiple topics. The user can do this semi-automatically (by requesting a split, and allowing the algorithm to determine where the best split is) or by manually selecting a set of terms from the InfoMagnet's word list and creating a new InfoMagnet using those words to represent the new InfoMagnet's topic. If the user finds words in an InfoMagnet's word list that lack topical relevance, the user can remove them from InfoMagnet's word list or from all the InfoMagnets' word lists at once.

Users may also choose to manually assign a segment to a topic by "snapping" that document to an InfoMagnet. "Snapping" is a way of overriding the attraction between the document and other InfoMagnets. By "snapping" a document to an InfoMagnet, the relationship between the "snapped" document and the associated InfoMagnet remains constant, regardless of any changes made to the InfoMagnet space subsequently.

If a user would like to remove the influence of a subset of the corpus from the behavior of the tool, the user may select an InfoMagnet and all the documents close to it and place them in the "quarantine" area of the interface. When placed in the quarantine, as when "snapped", a document's assignment remains unchanged. This feature is used to free screen space for the user.

If the user opts for segmenting each input discourse and working with topic segments rather than whole documents, an alternative interface allows the user to quickly browse through the corpus sequentially (Figure 2). By switching between this view and the bird's eye view, the user is able to see where each segment fits sequentially into the larger context of the discourse it was extracted from. The user can also use the sequential interface for making minor adjustments to topic segment boundaries and topic assignments where necessary. Once the user is satisfied with the topic representation in the space and the assignments of all documents to those topics, the tool can automatically generate an XML file, where all documents are tagged with their corresponding topic labels.

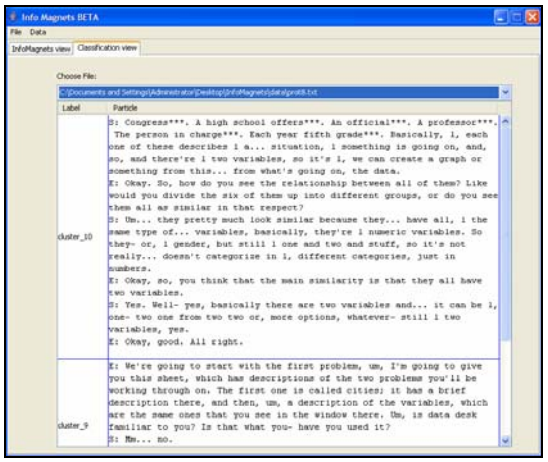

Figure 2. InfoMagnet's alternative sequential view

## 3 Implementation

As mentioned previously, InfoMagnets uses Latent Semantic Analysis (LSA) to relate documents to InfoMagnets. LSA is a dimensionality reduction technique that can be used to compute the semantic similarity between text spans of arbitrary size. For a more technical overview of LSA, we direct the reader to (Landauer et al., 1998).

The LSA space is constructed using the corpus that the user desires to organize, possibly augmented with some general purpose text (such as newsgroup data) to introduce more domain-general term associations. The parameters used in building the space are set by the user during pre-processing, so that the space is consistent with the semantic granularity the user is interested in capturing.

Because documents (or topic-segments) tend to cover more than one relevant topic, our clustering approach is based on what are determined heuristically to be the most important terms in the corpus, and not on whole documents. This higher granularity allows us to more precisely capture the topics discussed in the corpus by not imposing the assumption that documents are about a single topic. First, all terms that occur less than *n* times and in less than *m* documents are removed from consideration[2]. Then, the remaining terms are clustered via average-link clustering, using their LSA-based vector representations and using cosine-correlation as a vector similarity measure. Our clustering algorithm combines top-down clustering (Bisecting K-Means) and bottom-up clustering (Agglomerative Clustering) (Steinbach et al., 2000). This hybrid

[2] *n* and *m* are parameters set by the user.

clustering approach leverages the speed of bisecting K-means and the greedy search of agglomerative clustering, thus achieving a nice effectiveness versus efficiency balance.

Cluster centroids (InfoMagnets) and documents (or topic segments) are all treated as bag-of-words. Their vector-space representation is the sum of the LSA vectors of their constituent terms. When the user changes the topic-representation by removing or adding a term to an InfoMagnet, a new LSA vector is obtained by projecting the new bag-of-words onto the LSA space and re-computing the cosine correlation between all documents and the new topic.

## 4 An Example of Use

InfoMagnets was designed for easy usability by both computational linguistics and non-technical users. It has been successfully used by social psychologists working on on-line communities research as well as learning science researchers studying tutorial dialogue interactions (which we discuss in some detail here).

Using InfoMagnets, a thermodynamics domain expert constructed a topic analysis of a corpus of human tutoring dialogues collected during classroom study focusing on thermodynamics instruction (Rosé et al., 2005). Altogether each student's protocol was divided into between 10 and 25 segments such that the entire corpus was divided into approximately 379 topic segments altogether. Using InfoMagnets, the domain expert identified 15 distinct topics such that each student covered between 4 and 11 of these topics either once or multiple times throughout their interaction.

The topic analysis of the corpus gives us a way of quickly getting a sense of how tutors divided their instructional time between different topics of conversation. Based on this topic analysis of the human-tutoring corpus, the domain expert designed 12 dialogues, which were then implemented using a dialogue authoring environment called TuTalk (Gweon et al., 2005). In a recent very successful classroom evaluation, we observed the instructional effectiveness of these implemented tutorial dialogue agents, as measured by pre and post tests.

## Acknowledgments

This work was funded by Office of Naval Research, Cognitive and Neural Science Division, grant number N00014-05-1-0043.

## References

Jaime Arguello, Brian S. Butler, Lisa Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rose and Xiaoqing Wang (2006). Talk to Me: Foundations of Successful Individual-Group Interactions in Online Communities. *To appear in Proceedings of CHI: Human Factors in Computing*.

Matthew Chalmers and Paul Chitson (1992). Bead: Explorations in Information Visualization. *In Proceedings of ACM SIGIR*, 330-337

David Dubin (1995). Document Analysis for Visualization. *In Proceedings of ACM SIGIR*, 199-204.

Gahgene Gweon, Jaime Arguello, Carol Pai, Regan Carey, Zachary Zaiss, and Carolyn Rosé (2005). Towards a Prototyping Tool for Behavior Oriented Authoring of Conversational Interfaces, *Proceedings of the ACL Workshop on Educational Applications of NLP*.

Rohit Kumar, Carolyn Rosé, Vincent Aleven, Ana Iglesias, Allen Robinson (submitted). Evaluating the Effectiveness of Tutorial Dialogue Instruction in an Exploratory Learning Context, *Submitted to ITS '06*

Thomas Landauer, Peter W. Foltz, and Darrell Laham (1998). *Introduction to Latent Semantic Analysis*. Discourse Processes, 25, 259-284.

Anton Leuski and James Allan (2002). *Lighthouse: Showing the Way to Relevant Information*. In Proceedings of the IEEE InfoVis 2000

Matt Rasmussen and George Karypis (2004). *gCLUTO: An Interactive Clustering, Visualization, and Analysis System*. Technical Report # 04-021

Carolyn Rosé, Vincent Aleven, Regan Carey, Allen Robinson, and Chih Wu (2005). A First Evaluation of the Instructional Value of Negotiable Problem Solving Goals on the Exploratory Learning Continuum, *Proceedings of AI in Education '05*

Michael Steinbach, George Karypis, and Vipin Kuma (2000). A comparison of document clustering techniques. *In KDD Workshop on Text Mining*.

James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, and Anne Schur (1995). Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents. In Proceedings of IEEE InfoVis '95, 51-58.