Your Knowledge Partner™

The Use of Text Mining to Analyze Public Input

Josh Froelich, Megaputer Intelligence
Sergei Ananyan, Megaputer Intelligence
David L. Olson, University of Nebraska

Megaputer Intelligence, Inc.
120 West Seventh Street, Suite 310
Bloomington, IN 47404 USA
+1 812-330-0110

www.megaputer.com

Contents

Public Hearing Data

Public hearings are often held to gauge the pulse of the public with respect to important issues. These hearings provide useful platforms for individuals to express their opinions. Examples of such venues include traditional town hall meetings in New England, as well as public hearings conducted throughout the U.S. and in many other countries.

The results of these meetings are useful in that many ideas are expressed. In the past, much of the content has been lost due to the difficulty in recording what was said, the volume of comments, and the unstructured nature of the comments. Modern technology enables capturing this valuable information. Recording equipment can be used to gather all comments, which can be transcribed into digital form. Once a digital collection is formed, one can analyze this rich source of input using text mining software. The software enables identification of patterns and views of the support and opposition to various proposals.

This paper presents a case study of the use of text mining to evaluate citizen comments related to public issues. This is a highly unstructured domain, calling for an exploratory analysis. PolyAnalyst, a text mining software tool, was used to identify the importance of issues, to structure the comments into a meaningful form, and to develop support for various conclusions. The case demonstrates the potential of text mining to enable insight through keyword extraction, dimensional analysis, taxonomy classification, association analysis, and other useful tools.

The Hillsborough Planning Commission

The Hillsborough Planning Commission is responsible for transportation system planning in its community. As a governing body, the commission is required to evaluate and update their comprehensive plans every few years. Through a variety of meetings and surveys, comments were gathered from groups and individuals within the community. A large number of comments (850) were available for analysis (see Figure 1). After data was gathered, it was cleaned (placed in a proper format), and a number of initial variables identified. As with most text mining operations, structuring the data was an iterative process. Figure 1 shows the 420th record of these 850 comments. This individual expressed a desire to see a truck bypass to make it safer for golfers and provide better access of emergency vehicles to residential area. The data set consists of about 50 variables (13 shown in the lower left window), including identification and dating, agency jurisdiction, the recorded comment text, and a variety of categories selected by the commission which were filled out manually as the comment was entered into the database. These categories include the social service categories of mobility and transportation, education, social services, growth management and several others. This manual process of selecting which categories a comment belongs too by reading through every single one became extremely tedious. This is partly why the commission decided to involve a more automated analysis.

The Cleaned Dataset

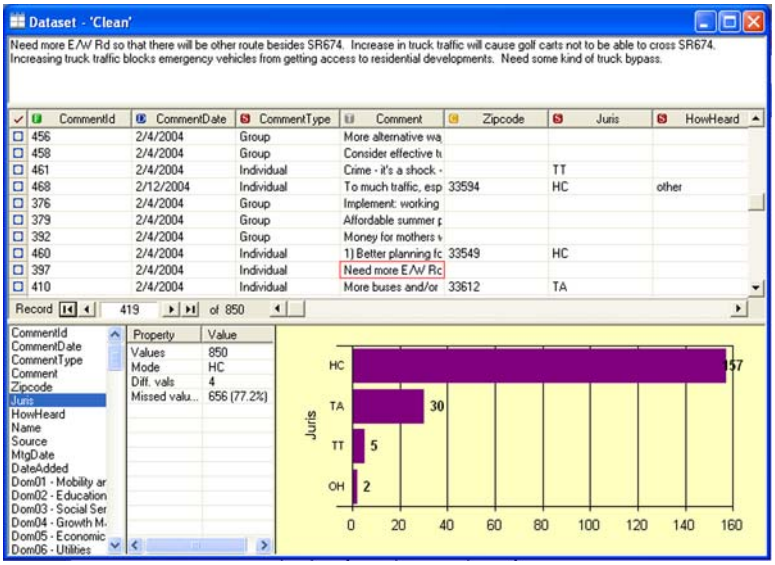

Figure 1: *Comment Data*

Initial Exploration

After cleaning and preparing the data for analysis, one of the initial goals is to gain a quick and summary understanding of concepts presented in each of the comments. One method used to ascertain this is by counting words and getting a list of the most frequent terms. This can be an iterative process and different parameters can be used to extract different types of keywords. Each word is examined individually, counted, and compared to other words.

Just counting words does not quite give a good picture of the main ideas. For example, a comment may contain the keyword “school” and the word “schools”. It is obvious to a person reading the comment that both words refer to the same idea, but not to a computer. A technique known as stemming is used here to identify alternative forms of different words and combine these different forms all into one word, or one idea. Yet this still does not accurately accomplish the task of identifying main ideas.

For example, a comment may contain the keyword “schools” and the word “education”. Again, a reader can easily identify that both words refer to the same idea, but a computer cannot. To solve this problem, PolyAnalyst incorporates a thesaurus of synonymous words. Through the use of a thesaurus, PolyAnalyst can identify that both “education” and “schools” should be counted together as the same idea. The dictionary also includes words with parent/child relationships, such as utilities (the parent) and children such as electricity, water, disposal, etc. Both synonyms and parent child relationships help the system go after the ideas behind the words, not just the words themselves. In practice this is referred to as semantic analysis, as the goal is to identify meanings.

A third dilemma in looking for key ideas is that many words that have little semantic value occur many times within the comments. These are words such as “the”,”a”, and “it”. These words are referred to as noise words, as they bias what are the most frequent ideas in the comments. The PolyAnalyst dictionary also incorporates a list of words to ignore in order to filter out these noisy words.

Figure 2 shows the results of semantic analysis of citizen comments. It lists key discovered terms specific to the context of the particular study.

Term	Record Count	% of Records
area	91	10.680751
community	44	5.164319
county	53	6.220657
development	78	9.154929
program	41	4.812206
road	87	10.211267
school	135	15.845070
service	77	9.037559
traffic	61	7.159624
water	45	5.281690

Figure 2: *Key Topics*

The word "area" appeared 110 times and in 92 records (different citizens comments), which is almost 11 percent of the total number of comments. The word "school" appeared even more, 180 times in 135 records. Figure 3 shows 10 records where the word "service" or its semantic variants appeared.

Rec #	service
719	We definitely need to increase transporation options - safe options, for people who can't or prefer not to use autos. We need to increase mass transit and provide safe pedestrian and non-motorized alternatives. This will reduce air pollution, help those who can't drive, increase healthy options and decrease congestion on the roads. We need a healthy mix of all transport modes. Thanks
57	colocate government services such as fire, sheriff, parks, schools, rescue services
285	More bus service - frequency, coverage, service hours
592	Our economic development strategies need to target high paying jobs not service industries, like tourist industries generate.
243	Expand the urban service area to include those areas of the urban expansion area which were removed 5 years ago - south to Boyette ELAPP tract and change RP2 back to R2 in that area. Extend Big Bend road to Balm Boyette Rd
471	Focus on infill development within the Urban Service Area
55	More services for seniors. Local community centers wanted to do things like painting and crafts.
446	Government needs to represent the public interests in the provision of private services(cable, telephone). Government needs to act on citizens complaints. Government needs to play a more proactive role on behalf of the public regarding public services.
562	Explore the possiblity of utilizing old CSX tracks for public transit to downtown
567	Improve HARTline service to USF

Figure 3: *Instances of the Word "Service"*

The software uses color coding to highlight words, making it easy for analysts to identify the presence of key terms. Notice that accessing "service" does more than literally identify the specific form of the word. In addition to "service", we can also see comments which contain the keyword "services". In addition, the dictionary allows the analysis to identify related terms such as "transit" (records 719 and 562).

Concept Identification

Once lists of key terms are generated, the analyst can gain a better picture of problems by category using a snake chart. Snake charts provide a means to display multiple dimensions against a small set of variables that can be color coded. A logarithmic function to emphasize outlying data is the metric for the snake chart plot (the more frequent the variable, the higher the measure). Figure 5 shows a snake chart of 14 selected variable topics and four example social service categories.

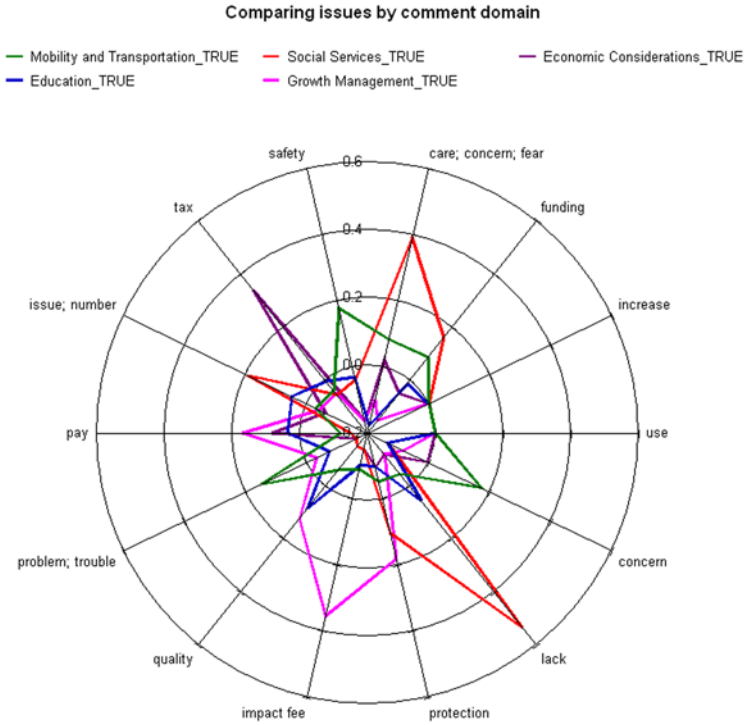

Figure 4: *Snake chart comparing K and J on various*

Key Word Analysis

Here the terms "impact fee," "use", and "pay" were associated with comments expressing concerns with growth management. Terms "safety," "problem or trouble," and "concern" were used most often with respect to comments about mobility and transportation. The terms "lack" and "care or concern or fear" were found most often with comments about social services. The software also allows the analyst to click on specific terms to drill down to obtain greater detail. For instance, Figure 5 shows a report of the results of clicking on the key term "lack" in Figure 4.

Rec #	lack
666	Lack of satellitee clinics/offices in centralized areas
554	lack of funding for social services across the board
553	Lack of human resoureces to perform more in-depth case management
520	Lack of adequate funding and community support for schools and teachers
487	How can we get the money to better finance education - lack of adequate funding

Figure 5: *Comments using "lack" under the "Social Services" category*

The analyst can now see details related to specific uses of the concept expressed by key words. The software also provides tools to show the distribution of comments across variables. Figure 6 shows the relative proportion of comments by the variable "jurisdiction," coded by issue domains (categories).

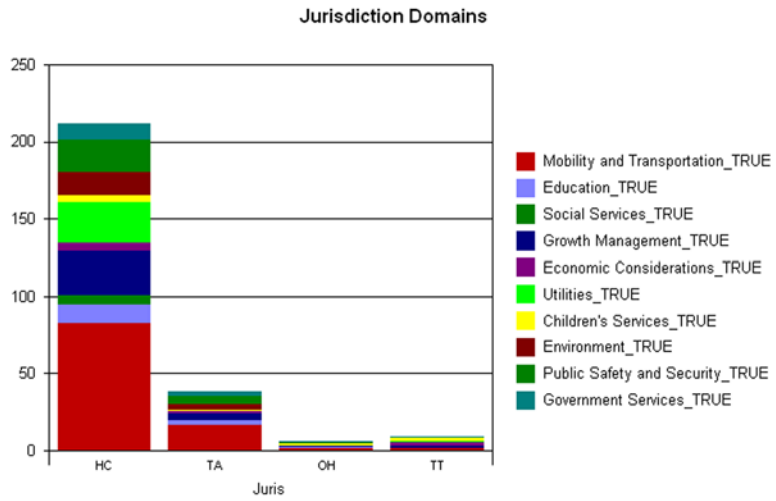

Figure 6: *Relative Distribution*

Here it can be seen that the vast majority of citizens interviewed came from Hillsborough Country (HC). A good proportion (about 80) of those naturally enough related to mobility and transportation issues. However, there also were about 20 comments relating to mobility and transportation pertaining to Tampa Area (TA).

Once mobility and transportation is identified as interesting, the analyst can focus on the relationships between different categories and different key words. Figure 7 shows such an analysis.

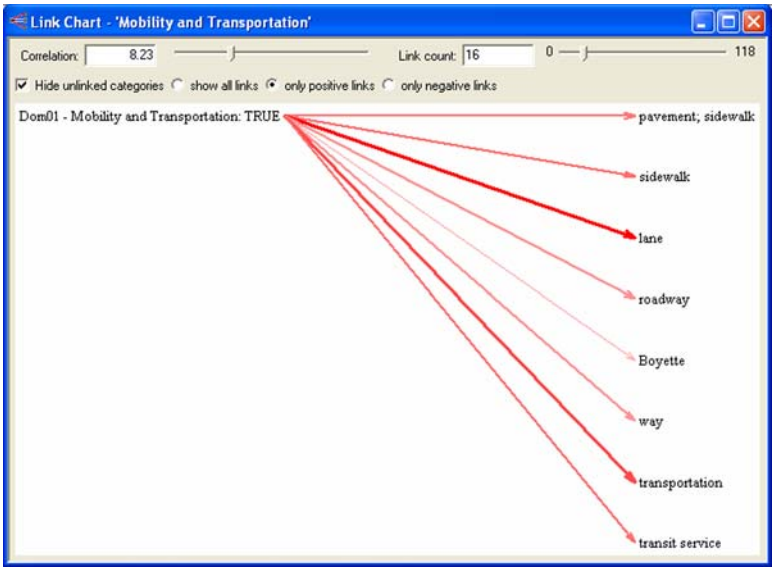

Figure 7: *Keyword Relationship to Variable Mobility and Transportation*

The heavier lines indicate stronger relationships. In this case, school was found to be highly correlated with comments that did not involve mobility and transportation. On the other hand, "road," "traffic," and "lane" did correlate strongly with mobility and transportation. Keyword relationship analysis can be used as a test to determine whether the coded domains match the discovered keywords. If there is a good match, this indicates that the prior work of identifying keywords was accurate. If there were not, the analyst could reiterate and identify more suitable key words.

Modeling

Another useful tool is a decision tree model. This form of model generates if-then rules based upon the proportion of relationship among variables and outcomes. Figure 8 shows a decision tree for this example.

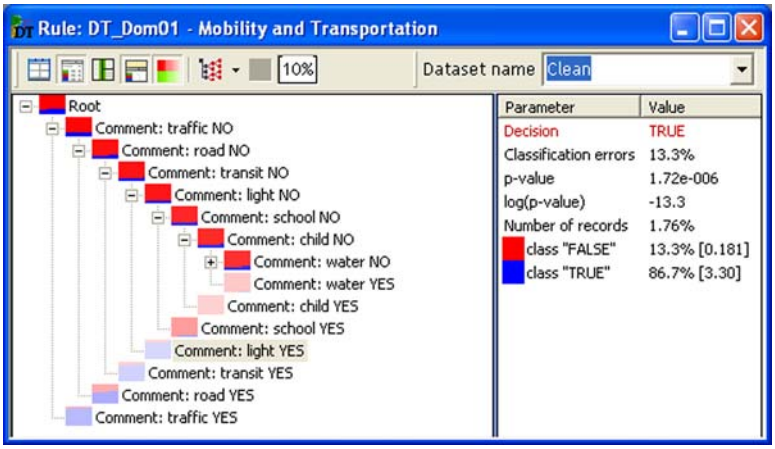

Figure 8: *Decision Tree for Transportation Model*

If the set of keywords "traffic," "road," "transit" and "light" were present, there was a high likelihood (0.867) that the comment relates to mobility and transportation issues. This model was based on 1.76 percent of the 850 cases in the data set. The decision tree model in this case identifies useful combinations of terms that predict the presence of other key terms. Should new comments arrive containing these four terms and where the category is not already know, it is highly likely that it relates to mobility and transportation issues.

Another useful modeling tool is link analysis. Link analysis graphically displays the relationship among variables (see Figure 9). The sliders at the top of the window give the analyst the ability to control how many relationships are displayed. The sliders filter out relationships based on correlation. If the left slider sets a low minimum, most if not all of the variables will appear, with arrows connected to just about every other variable. This is too cluttered a display, so the analyst can slide the minimum up until an appropriate number of relationships appear. If the slider is raised too high, the screen will probably be blank. The right slider is available for the maximum value. Variables are color coded as selected by the analyst. Here icons are used for each of the seven issue areas displayed (such as an open book for educational issues). Arrowed lines show which variables have the specified degree of relationship. Different levels of relationship can be indicated by darker arrows.

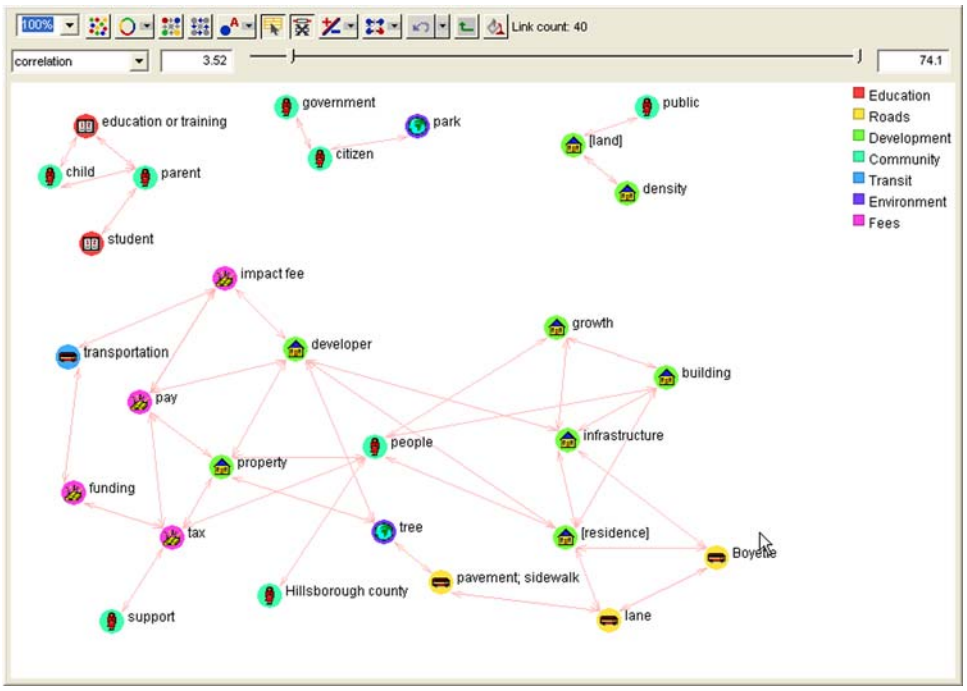

Figure 9: *Link analysis of Keywords*

In this case, two environmental issues appear. System and infrastructure key terms appeared with the environmental issue of water, while the environmental issue relating to trees appeared related to the keywords developer, property, and pavement or sidewalk. The link analysis chart helps the analyst see which problems are most associated with which issues. By clicking on specific arcs, the system will display those comments involved. This is shown in Figure 10 for the link between transportation and funding.

Comment
Now that the (community)plan is nearing completion, plans need to made for near term implementation with funding appropiated as neccessary. This applies to all aspects of plans including TRANSPORTATION CORRIDOR PLAN.
I am concerned about the lack of progress by the County Commissioners to identify sufficient funding for transportation (roadway) improvements. Reallocation of resources within current budgets has failed to make significant headway in fixing problems now. The Commissioners need to step up and propse tax and/or impact fee increases. If necessary let the voters decide what they can bear in a referendum.
Need regional-level funding for regional-level transportation needs. Impact fees possible?
More funding for alternative transportation choices
Specifically concerning the southshore area our most urgent transportation need is to improve (widen and lights) US301 from Riverview to SR674. Concurrently we need to get ahead of the growth curve by aggressively implementing the remainder of the southshore corridor plan. Countywide remains the issue of properly funding all our transportation plans or they become useless

Figure 10: *Comments Related to Selected Link*

This allows the analyst to see specifics about the relationship that was identified. The comments here express what kind of changes they would like to see in transportation funding.

Text Categorization

The knowledge gained by the analysis to date can then be used to more thoroughly explore keywords by issue. Text categorization is demonstrated in Figure 11, where suggested groups of comments are created by issue.

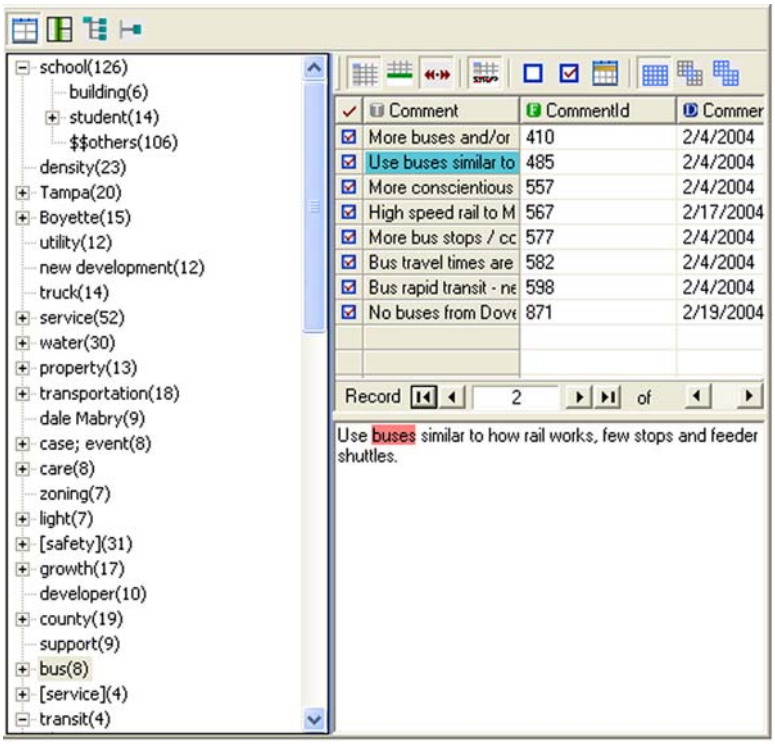

Figure 11: *Key Kategories*

The left window gives the entire taxonomy. There were 126 instances of the keyword "school." There were 8 comments that related to the keyword "bus" and its related terms. These are all displayed in the upper right window. Record 421 is selected, and the complete comment given in the lower right window. Those keywords with the "+" symbol on the left can be clicked on to obtain subcategories.

Here the system generated eight example groups of comments created by the text categorizer:

1. Education
2. Roads
3. Service
4. Water
5. Property
6. Transportation
7. Safety
8. Buses

The analyst has the ability to override this suggested list of groupings.

The use of keyword analysis, link analysis, and automatic categorization help the analyst to gain an immediate and comprehensive understanding of the main ideas in the comments. With this understanding, the analyst can perform more directed analyses providing evidence for making certain conclusions.

Dimension Analysis

The next phase of an exploratory text mining analysis could create dimensions for more complete analysis. Figure 12 shows an example where six variables are displayed in an on-line analytic processing (OLAP) form.

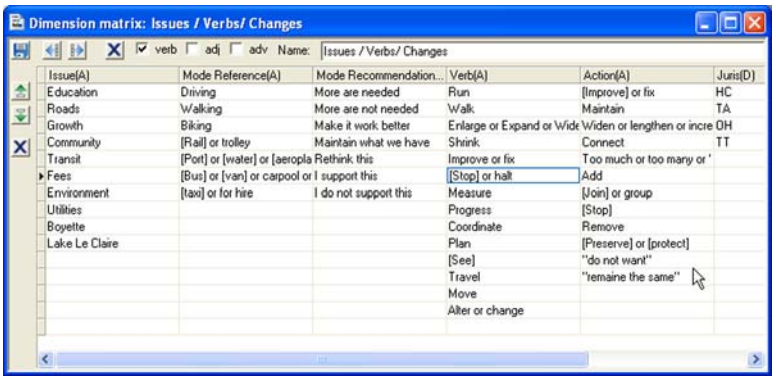

Figure 12: *Definitions of Dimensions for the Analysis*

In this case, two types of variables are displayed, category fields and text fields. For instance, the "How Heard" dimension is based on the categorical variable that was entered for each comment. Each row of the "Issue" dimension is a search query which matches certain comments based on matching keywords. This matrix can be used to create dimensions such as "key citizen concerns" and then compared to other dimensions, such as those comments willing to pay for resolution as opposed to those who would not be willing to pay.

The analysis can be carried further by additional structuring, as demonstrated in Figure 13.

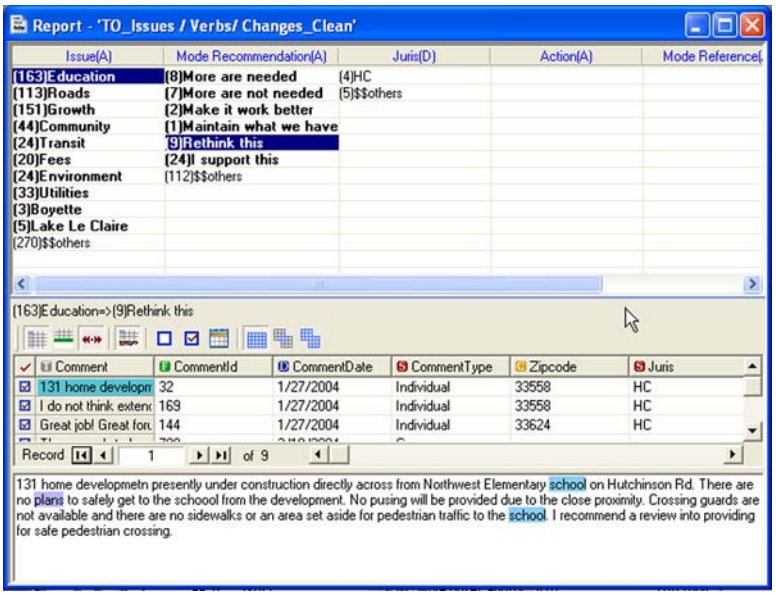

Figure 13: *OLAP Report*

The analyst is interested in educational issues (for which there were 163 comments). Within those 163 comments, 9 expressed the need to "Rethink this." None of these 9 comments were obtained from the five given "How Heard" sources. The first of these records was record number 20, made on 1/27/2004 by an individual from zip code 33558. The comment is given in full in the bottom window. All keywords are color-coded for cross-identification.

The commission might be interested in comments related to the counties road infrastructure. Figure 14 shows the OLAP report of those comments related to road widening. Note that of these 16 comments, 8 were in favor of additional widening, 2 gave indication of support, and 2 opposed the idea. The other 4 comments did not express a pro or con position. Any of the specific comments can be selected to allow the analyst to drill down and read the specific comment.

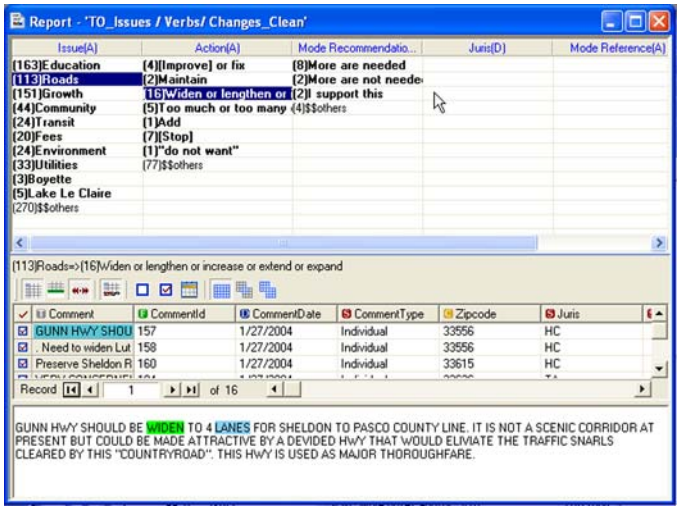

Figure 14: *Widening Comments*

Figure 15 gives opposing views expressed by the keyword "Stop." There were 7 of these comments, each expressing the opinion that "more are not needed." Specific reasons can be quickly identified by selecting specific comments.

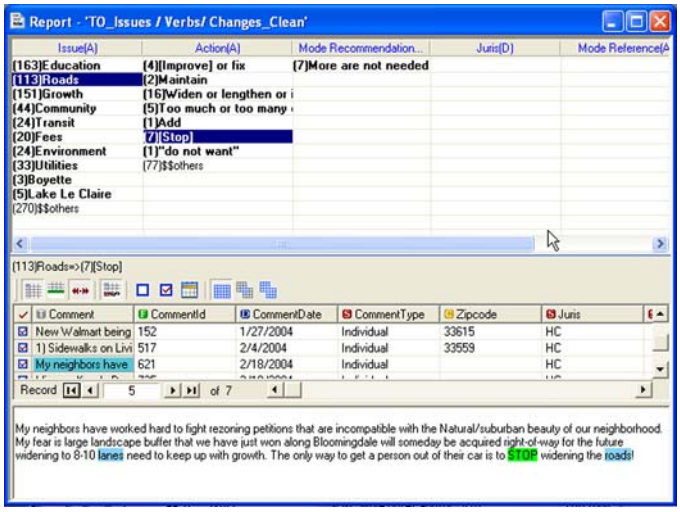

Figure 15: *Not to Widen*

Final taxonomies can be used to sort out the relative support for various sides of any issue. Figure 16 shows a taxonomy of issues on two dimensions. This splits up the comments on selected categories chosen and defined by the analyst. Here the first level matches different issues. The second level matches relative pro and con support or sub categories.

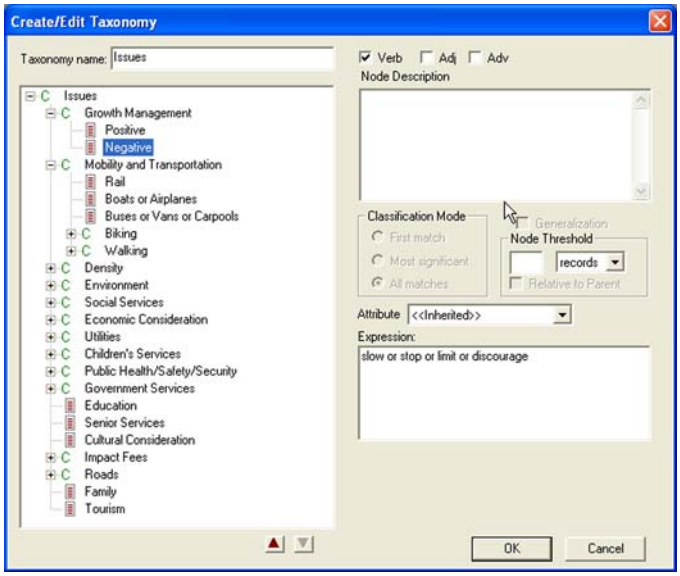

Figure 16: *Taxonomy of Issues*

To demonstrate the taxonomy in action, Figure 17 gives a drill-down analysis of those positive about the environment category. By selecting this categorical node from the taxonomy on the left window, a list of the six matching comments is obtained in the upper right window. The fifth of these is selected, and the full text of the comment given in the lower right window. Keywords are color coded for easy analyst identification.

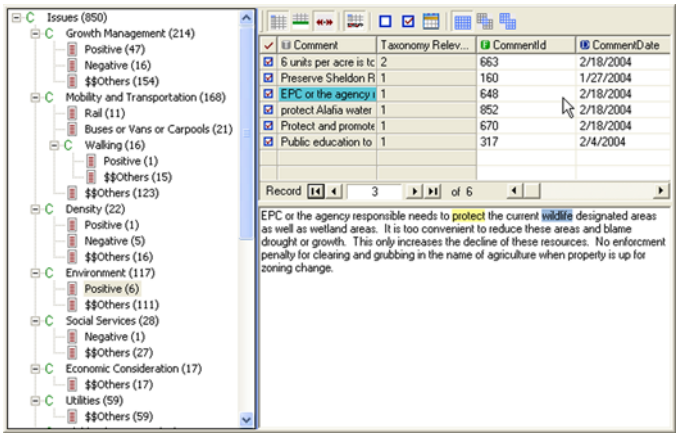

Figure 17: Those Supporting the Environment

Conclusions

Text mining through PolyAnalyst software provides a tool to quickly find key topics in unstructured data environments. Supporting tools give a means to discover links between topics, such as family values and education. The software enables discovery of comment categories, and generation of issue dimensions and taxonomies that enable monitoring comments for issues of interest and better overall understanding of issues, their distributions, and measures of concern to the public. Text mining can take unstructured data and process it to lead to greater understanding. This is especially important when dealing with public issues, where the arguments for and against particular positions are important to identify and calculate, but are stored in natural language comments. Text mining offers a valuable tool to support the process of public input analysis, and knowledge discovery and reporting.

Corporate and Americas Headquarters
Megaputer Intelligence Inc.
120 West Seventh Street, Suite 310
Bloomington, IN 47404
TEL **+1.812.330.0110**; FAX **+1.812.330.0150**
EMAIL info@megaputer.com

Europe Headquarters
Megaputer Intelligence Ltd.
B. Tatarskaja 38
Moscow 113184 Russia
TEL **+7.095.951.8079**; FAX **+7.095.953.5731**
EMAIL info@megaputer.com