Employee Survey Analysis

Josh Froelich, Megaputer Intelligence
Sergei Ananyan, Megaputer Intelligence

Contents

Text mining tools provide the ability to take unstructured data in the form of digitized text, and translate this data into useful knowledge. Semantic text analysis combines quantitative computer support with expert experience to process text data qualitatively. This can give new views of data that can lead to human insight. This paper demonstrates the process by which PolyAnalyst can take employee survey data and develop business taxonomies showing the logical organization of responses, making it easier to see the views held on each issue along with their relative support.

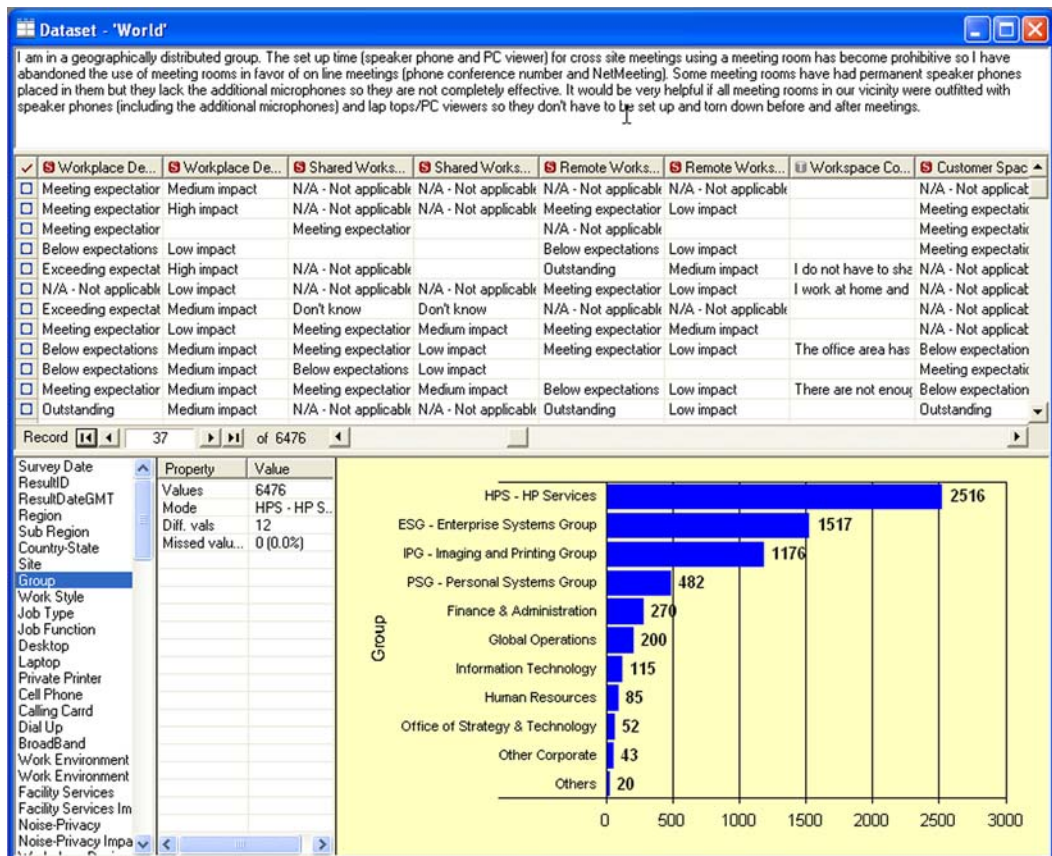The data context in this case is employee survey data, an extract of which is shown in Figure 1.

Figure 1: *Employee Survey Data*

An example survey response is shown in the top window. In this case, the respondent is explaining why meeting room technology was not used, pointing out the simple enhancements needed to make meeting room technology more effective. The dataset includes 105 attributes reflecting profile characteristics of respondents, demographic data, time data, and both structured and open-ended responses for 6,476 entries. One of these attributes, "Group," is displayed in a histogram. About 2,500 of the responses came from HP Services, for example. The window on the lower left indicates some of these 105 attributes. Structured responses can be dealt with by traditional data mining techniques. There is a great deal of information, however, available in the unstructured text data. PolyAnalyst software tools provide the ability to tap into this useful source of knowledge.

Semantic Text Analysis

This paper's aim is to demonstrate the process through which unstructured text data is converted to useful knowledge. PolyAnalyst Text Mining software includes user dictionaries, which can be edited for each specific study. Figure 2 shows the primary window to access a user dictionary.
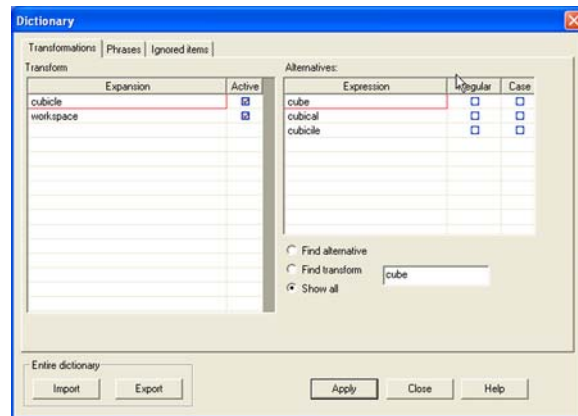

Figure 2: *User Dictionary*

An important step is to enter items to be ignored, eliminating common terms that are not interesting in this particular study. The semantic dictionary keeps track of important terms in a variety of forms, both by different forms of root words, as well as synonyms. A basic semantic lexicon is part of the system, but the user can enhance this resource for particular studies. Once dictionary settings are made, the system can count the number of occurrences of key words. Figure 3 shows text analysis results for workspace comments in the employee survey database.

Rule name	Rec Count	%
Workspace Comment_noise	335	50.99
Workspace Comment_workspace	163	24.81
Workspace Comment_office	156	23.74
Workspace Comment_people	144	21.92
Workspace Comment_area	103	15.68
Workspace Comment_privacy	131	19.94
Workspace Comment_phone	136	20.7
Workspace Comment_noise level	74	11.26
Workspace Comment_desk	75	11.42
Workspace Comment_meeting room	76	11.57
Workspace Comment_space	95	14.46
Workspace Comment_environment	116	17.66
Workspace Comment_meeting	92	14
Workspace Comment_room	126	19.18
Workspace Comment_conversation	72	10.96
Workspace Comment_due	47	7.154
Workspace Comment_cubicle	42	6.393
Workspace Comment_customer	39	5.936
Workspace Comment_meeting call	37	5.632

Figure 3: *Text Analysis Results*

It can be seen that workspace comments included semantic variants of noise in 335 records, almost one-half of the almost 700 records with workspace comments. This list guides the analyst to identify workspace features that triggered comment by those surveyed. Some comments addressed workspace (or office) in general. The next most common term related to people. Specifics for any particular phrase among workspace comments can be selected, allowing the user to drill down to more details. Those comments relating to workspace distractions can be gathered in a subset of the data. There were 163 such records. The analyst labeled this data subset as "Workspace Distraction." Figure 4 shows drill-down results for the phrase "workspace" in the workspace comment attribute.
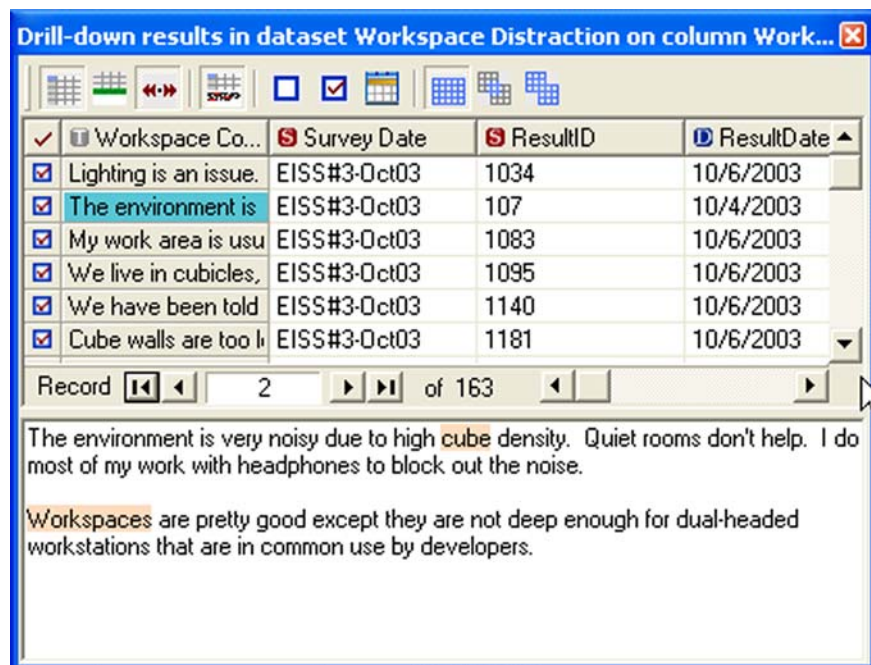
Figure 4: Distraction Drill-Down Results

The second line in Figure 4 relates to a comment made in October 2003 relating to noise in the environment, blaming the high number of cubicles in the room. However, this employee has coped with the noise problem through the use of headphones. Thus, the conclusion of this subject is that workspace would be acceptable, given sufficient depth to allow dual-headed workstations. The ability to drill-down enables the analyst to make more sense of what subjects are trying to say relative to noise or any other workspace comment key word.

Quantitative Models

PolyAnalyst includes a number of data mining tools that can be used to quantitatively analyze data. For instance Figure 1 above showed a distribution chart. Figure 5 shows a distribution chart for the attribute "Work Environment," which includes 5 possible values, as well as "Don't know" and "Not applicable."
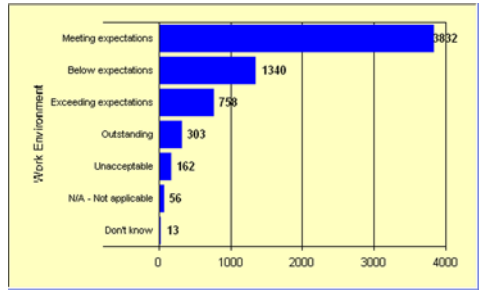
Figure 5: *Distribution Chart*

The distribution chart provides a quick view of the number of responses for each attribute value. Most survey responses were satisfied with current conditions, but about 1,250 felt that the work environment was below expectations. Work environment opinions in this database can be evaluated longitudinally, as survey data is available over the period October 2003 to February 2004. Figure 6 shows trends for the attribute "Work Environment" over time.
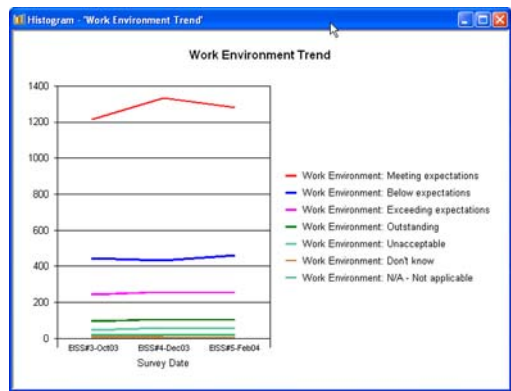

Figure 6: *Trend Chart*

Link charts enable analysts to view the correlation among attribute values. Figure 7 shows those combinations of attribute values for "Region" and "Work Environment" that have strong correlations. The boldness of the arcs linking attribute value also provides a representation of strength of correlation.
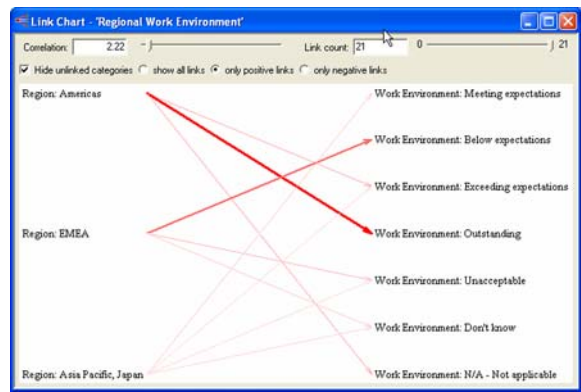
Figure 7: *Link Chart Across Attributes*

The PolyAnalyst software includes the ability to generate decision trees of association rules. Figure 8 shows such a decision tree.
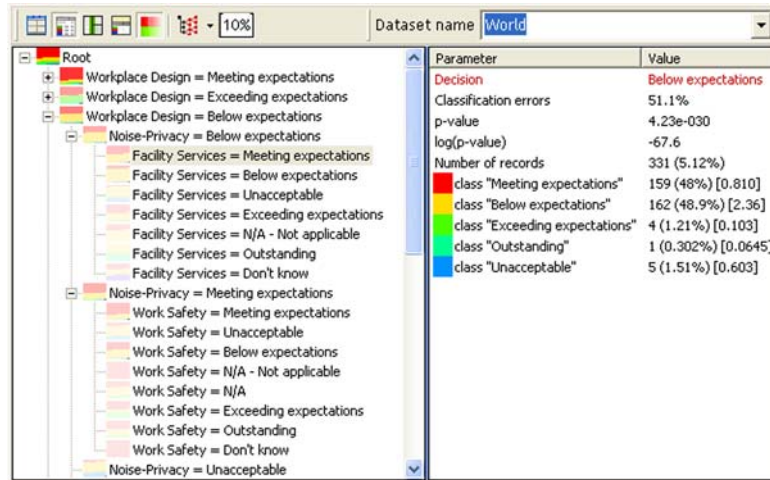
Figure 8: *Decision Tree of Work Environment*

In this case, if "Workplace Design" has an attribute value of "Below expectations," and noise-privacy has a value of "Below expectations," there were 333 records (5.12% of the 6,476 total), among which 1.51 percent were rated as unacceptable on attribute "Facility Services," 48.9 percent were below expectations, 48 percent met expectations, 1.21 percent exceeded expectations, and 0.302 percent were rated as outstanding. The analyst here selected the subgroup with the Facility Services rating "Meeting expectations," and those records are displayed in the lower right window. The user can view these approximately 160 records in detail. In general, however, the most common rating when workplace design and noise-privacy were below expectations was that facility services were also below expectations.

Unstructured Text Analysis

Text mining software's value in dealing with unstructured text can be demonstrated here. First, keywords are identified. Figure 9 shows the count of records with those keywords displayed.

Figure 9: *Keywords*

Of those keywords displayed, "office" showed up in 438 records (6.76 percent of all 6,476 records). The software can display a tree of these keywords, as exhibited in the left window of Figure 10.
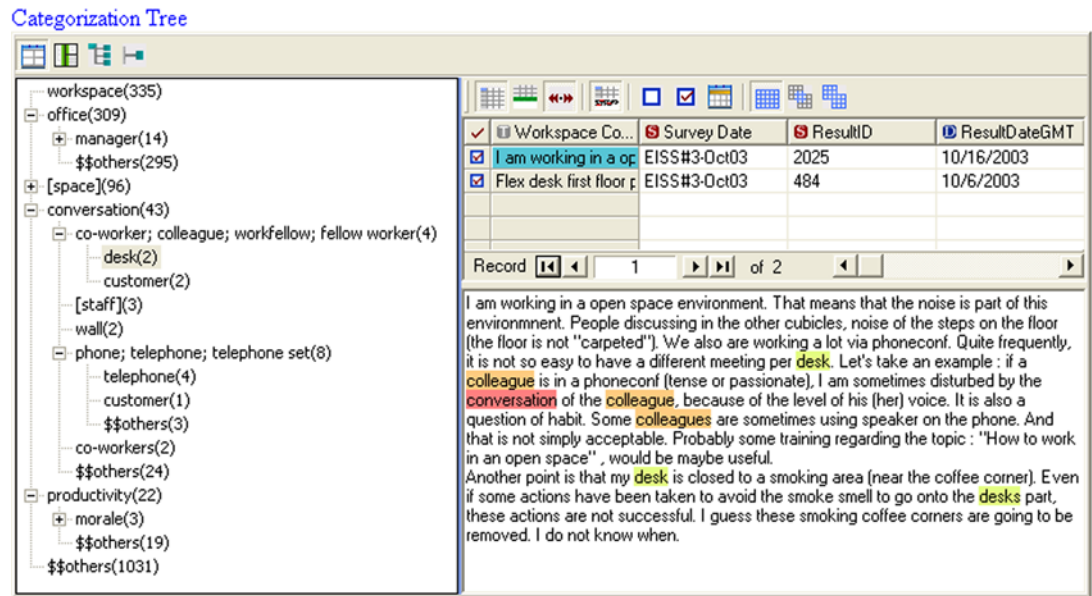
Figure 10 : *Category Development*

Here there were 43 occurrences of the key word "conversation, four of which involved co-workers or its synonyms. Two of these contained the key word "desk," which was selected by the analyst. These two records are shown in the upper right window. The first of these is selected, and the lower right window displays the survey response, with key words highlighted in coded color. Figure 11 shows further drill-down results for six records of 389 relating to key word "workspace."

Figure 11 : *Drill-Down on Key Word*

This also is color coded to quickly guide the user to key terms.

Taxonomy Classification

A taxonomy classification can be developed through a general process. Those terms important to a specific study can be identified, based upon survey responses. In Figure 12, the left window displays a set of well-defined categories used to group terms defining subsets.
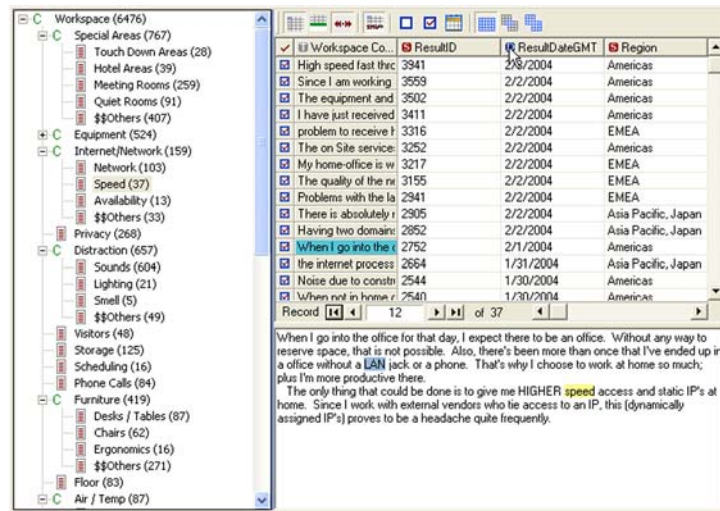

Figure 12: *Taxonomy Classification*

The analyst has selected the term "Speed," a subset of workspace comments involving "Internet/Network." There are a total of 37 such records, the first 11 of which are displayed in the upper left window. The first, record number 268 is selected, and the full comment displayed in the lower right window. Key terms are highlighted by color. Development of this taxonomy can guide the analyst to identify key issues.

This process can focus on key issues. For instance, 657 of the total 6,476 comments involved workspace distractions. These 657 records were exported for detailed analysis. Figure 13 displays the first three of these comments.
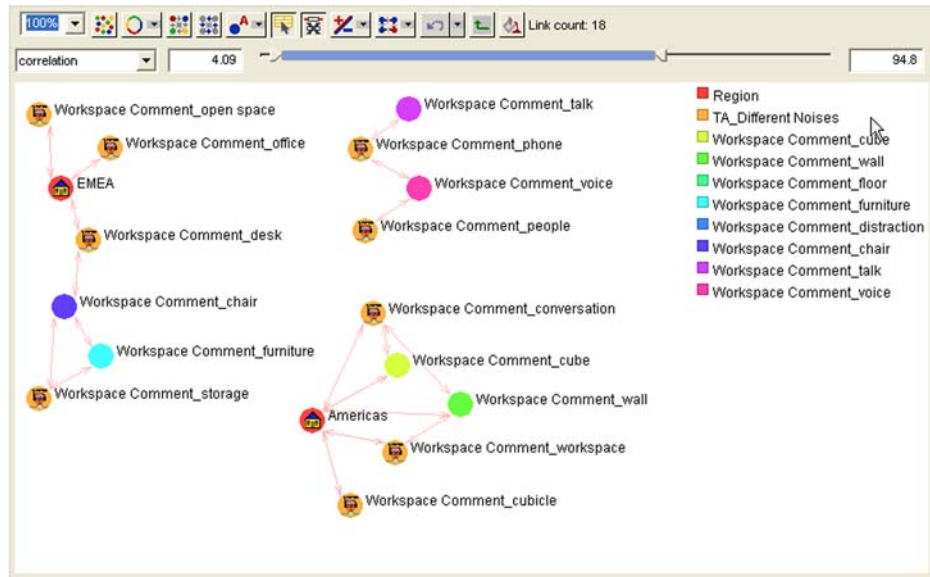

Figure 13: *Export of Records Focusing on Distractions*

Noise is apparent in these three comments. Other distractions identified from other records include poor lighting, and bad smell. Analysis can look at other factors related to distractions. Figure 14 shows a link chart showing key terms related to workspace comments.

Figure 14: *Link Chart of Distraction Comments*

This link chart is also color coded, with key on the upper right. Two regions show up. In the Americas, comments relate to noise in cubicles, conversations, and workspace in general. Other strong correlations show up for cube and walls. In area EMEA, comments relate to noise, associated with desks, offices, and open space. Figure 15 shows key correlations among this subset of data.
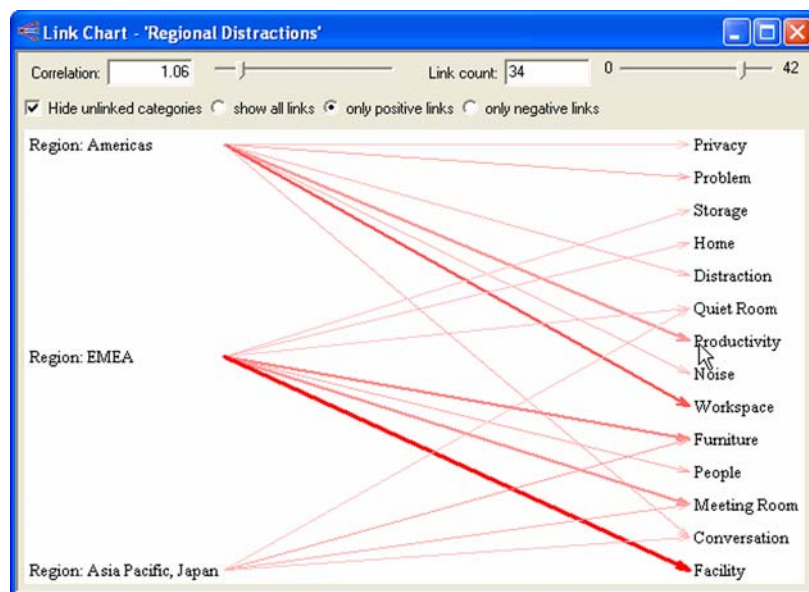
Figure 15: *Correlation Links*

Strongest correlations displayed in the Americas are with the workspace, and productivity. Strongest correlations in the EMEA region are with facilities, furniture, and meeting rooms. Figure 16 shows an additional quantitative tool, a snake chart.
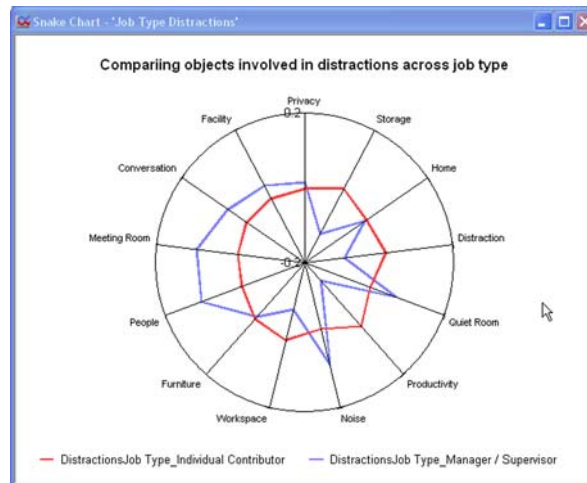
Figure 16: *Snake Chart*

The snake chart displays correlations visually on multiple (here 13) dimensions, with stronger correlations graphed at greater distance from the center. Two attributes were selected for display on this snake chart (distractions reported by individuals, and distractions reported by managers). Managers had fairly equal correlations with all thirteen distractions graphed. Individuals had lower correlations with storage and workspace, and less reported impact on productivity and distractions than did managers). The analyst can drill down to reports such as shown in Figure 17.
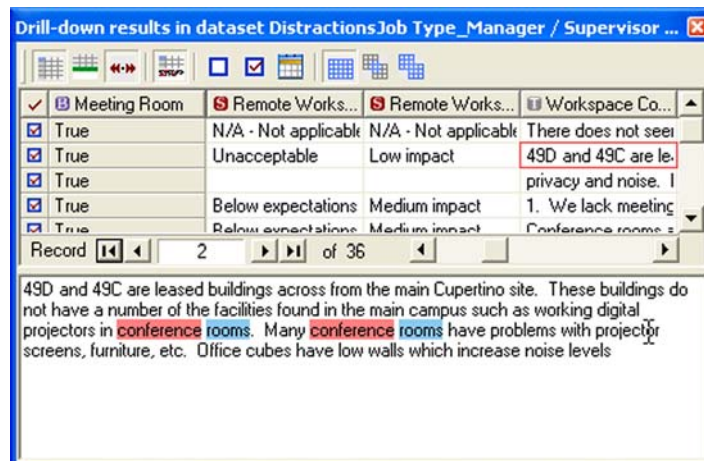
Figure 17: *Drill-Down Report for Distractions by Managers*

The snake chart, supplemented by specific reports obtained by drilling down, indicates that managers complain about difficulties in meetings, while individual employees complain more about lost productivity due to noise.

OLAP Charts

The text analysis of generated key terms can then be used to sort out the data by selected attributes. This is demonstrated in Figure 18, which shows who is affected by noise, how they are affected, and where.
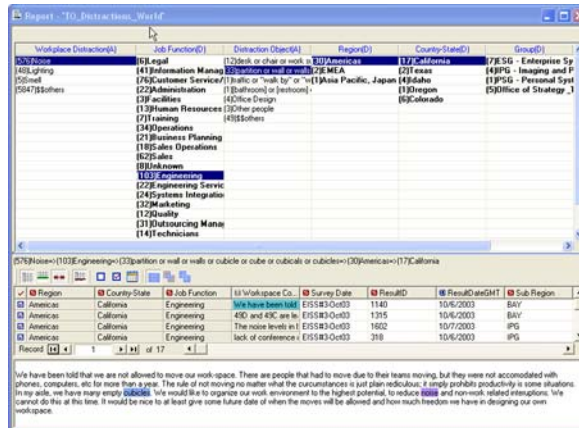

Figure 18: OLAP Chart

The analyst can select the header attributes. In this case, of the 6,476 comments in the database, 576 related to noise (48 to lighting, 5 to smell). The analyst here has selected "Noise" for further analysis by job function. Of the 576 comments relating to noise, 103 involved engineers. Of those 103, 33 involved partitions or walls. Of those 33, 30 were from the Americas region, 2 from the EMEA region, and 1 from the Pacific. Of the 30 from the Americas region, 17 involved California. The analyst could search by group, but here those 17 records are available for detailed review. The first is displayed in the bottom window, with key terms highlighted by color.

In this case, the data indicates that engineers in California want better cubicle dividers to reduce noise. This is a demonstration of discovered knowledge, something that the user or analyst would not have known to expect without going through the process of text analysis.

Summary

Text mining using PolyAnalyst software is supported by quantitative and qualitative tools. Quantitative support begins with visualization, providing histograms, pie charts, bar charts, and snake diagrams to show the relative density of key terms. Other quantitative support is provided by correlation models, through link charts, and decision trees sorting key terms. Qualitative support is provided by the process of identifying keywords, which are categorized by the software, and a series of reports of subsets of data selected by the analyst. This leads to development of business taxonomies, which can lead to better understanding of survey data than would be possible by traditional methods where all variables have to be set up prior to analysis. In this example, focus on the concept of distraction led to identification by region and job category. Analysis of the OLAP dimension matrix revealed support for issues such the inadequate cubicle walls, mostly affecting engineering and customer service staff.

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