Architecture for Integrated and Dynamic Data Analysis (AIDA) Client Guide

Contents

Overview

The Architecture for Integrated and Dynamic Data Analysis (AIDA) is an information analysis tool initially developed by Argonne National Laboratory and the University of Chicago. The tool allows analysts to view information from many data streams in a way that can alert them to or increase their situational awareness of potential and on-going events. More recently, the University of Alaska and University of Alaska Anchorage have applied and modified this tool to focus specifically on identifying and understanding land use events and land use change. In AIDA, analysts can view and assess data in order to identify situations in which significant events may occur. AIDA facilitates the identification of potential events by providing information on relevant term linkages. Throughout this document, examples of AIDA will focus on a case study involving land use issues and events in Iowa and Nebraska.

Client Guide Purpose

The purpose of this guide is to demonstrate how to use the client component of AIDA and evaluate results returned by the server engine. This document will not focus on specific coding issues, such as the Java classes implemented to allow the functionality discussed. However, users can look at the Javadocs in the /Client/docs folder for information about classes used.

Client Guide Goals

When completing this guide, users will be proficient in, 1) selecting keywords for searchs, 2) defining a visualization method for search results, 3) exploring search results in different visualization types, 4) searching for terms within the returned document cache, 5) evaluating search results, 6) locating and having access to text output for further statistical analysis.

AIDA Client Purpose

The purpose of the AIDA Client is to provide users with a tool that searches for trends within and links between data from a variety of sources. With this information, analysts can identify early, potentially significant events. Key AIDA features include:

- User defined keyword search capability
- Keyword semantic mapping
- Keyword time series visualization
- Document search

Information Fusion

Information fusion, as applied in AIDA, is the process through which data are integrated according to semantic content. When the process is complete, associations between documents from a variety of sources are made, and those with shared terms are linked. For documents with the same primary subject, this can be used to count occurrences for

trend analysis across the time scale investigated. For documents with different primary subjects but shared secondary terms or co-occurrences, a link is identified based on the shared terminology. In such an instance, the trend in usage of the primary term and the link with the secondary term are used in creating a semantic map.

How AIDA Supports Land Use Issues

In order to respond adequately to significant land use issues that may threaten or affect social wellbeing, a robust surveillance tool is required in order to find factors relevant for land use change, particularly such issues that are not easily observed using other monitoring devices (e.g., satellite data). Potentially, there are many text data and information sources available to stakeholders. Finding an effective means of using the information received can be difficult due to the volume of information. AIDA is a system through which data can be categorized and parsed for links and meaning. Some of the ways AIDA can be useful are:

- Identification of trends in reported use of single keywords. Changes in occurrence relative to the mean value of keywords related to land use terms could be evidence of a growing focus on specific issues.
- After identifying a behavioral anomaly within the data, users may drill down into the portion of the cache in which the change occurs. The search function enables the analyst to explore data for links based on other terms not necessarily related to the primary term search. An example of this would be to search within the data on the term *harvest* at a given time slice for terms that might indicate a relevant story to the primary term (e.g., *legislation*, *stimulus*, *policy*, *weather*, etc.).
- During an event, inflection points within a trend analysis of unstructured data can prove instructive in understanding what information is being conveyed to the public from media outlets.
- Links between land use keywords (e.g., *agriculture*, *urban expansion*) could identify a growing trend in changes affecting land use.
- Provide statistical output for conducting more quantitative analysis (e.g., principal component analysis) that can identify linkages and associations between terms that may inform on land use trends.

AIDA Client Functions

The features included in the AIDA Client include keyword selection, keyword visualization, data exploration, 3D semantic map, 2D semantic map, 2D time series, 2D structured data charts, document search, and GIS. However, this document only focuses on unstructured data use; therefore, structured data charts and GIS use will not be discussed.

Components

Cache: The cache is a collection of data that has been received and parsed by the AIDA Core component. The Iowa and Nebraska caches (/Client/cache_Iowa and /Client/cache_Nebraska) are drawn from newspapers from those states and represent the analysis conducted in the AIDA Server. The dates covered by those caches are 03/01/09-11/01/09. Users can interact with these data through the tools described below.

Keyword Selection: In the Iowa and Nebraska examples, the current keyword selection box allows users to choose from five lists of pre-loaded terms before running the data visualization tool. Three of the options are dictionaries related to areas of importance to major land use issues: "Agriculture," "Urban and Transportation," and "Forestry and Grassland." These are defined based on a preliminary search of newspaper sources, which indicated the relative importance of these topics, and their associated terms, on land use issues. Other land use issues and terms that may have indirect relevance (e.g., *legislation*, *policy*) are included in an "Other" category. The user may also search on all keywords concurrently. The AIDA system is designed to allow for expansion of the pre-loaded keyword lists as users determine new information or alternative expert databases are used.

Unstructured Data Visualization: Based on the users' needs, they can select from one of three ways of visualizing unstructured data, semantic mapping in two and three dimensions, and in graphing data in a time series. Each can inform the user of how to best proceed in refining their search area and defining terms of interest when switching between tools.

- **2D Time Series:** The 2D time series data visualization consists of two components, a line graph divided into time slices of a running count of documents that include a keyword. Below the graph is a dynamic legend with which the user can sort and refine what is displayed in the graph.

- **Semantic Maps:** Semantic mapping is a method of visually organizing the relationships between data. For the purposes of the AIDA client and land use example, the information is keywords related to land use issues in Iowa and Nebraska that are displayed as nodes in the graphing space using an icon, in the case of AIDA a yellow disk. The disks vary in size based on the number of times the keyword occurs in the time slice, relative to the mean number of times it occurs in the cache up to that point in time. Lines connect terms in instances of their co-occurrence within or between documents.

    - **2D Semantic Map:** The two-dimensional semantic map visualization displays a map of the selected keywords within a single time slice of the cache. The user can move between time slices and see changes in keyword frequency and links between keywords. Analysts also have the capability to explore the data within a node and link to the original data source through a documents list.

- **3D Semantic Map:** The three-dimensional semantic map displays virtually the same data as the two dimensional version. The key difference is that in the 3D visualization the user is able to see not only the mapping of terms relative to past values and the links between them, but they can also see the other time slices. This gives a better visual assessment for changes in individual nodes (indicating trends) and the strength of linkages.

Text Output: In addition to visualization, statistical output of node term counts, link counts, and number of documents per time slice are produced (see /Client/text_cache_output) when applying either the 2D or 3D semantic maps. The output data produced reflect the terms chosen and removes previous data in the text_cache_output folder. The files produced start with "Nodes" (for node counts), "Links" (for link counts), and "Summary" (for document counts); these three file types are made for each time slice, with the time interval used to create the full file titles. Output can be exported into R (http://www.r-project.org/) or other statistical tools for further analysis.

Document Exploration: AIDA gives users the ability to explore within the data that is used to generate the semantic map visualizations. Selecting nodes and links within those visualizations opens a window in which document headlines, very short summaries, and links to the original sources are located.

Document Search: The Document search function allows users to search within a cache for selected documents containing ad hoc terms not included in the keywords dictionaries.

Using the AIDA Client

The AIDA Client is a dialog between the user and the cache of information, with the software acting as an intermediary. Analysts can sift through as much or as little of the information as they desire. The user defines what information he or she wants to see based on keywords and how that information should be displayed. Figure 1 is a workflow diagram for a typical use of the AIDA Client. The first three steps involve user selection and are consistent no matter what visualization scheme the user chooses. Once the visualization selection is made in the fourth step, the workflow diverges based on the selection.

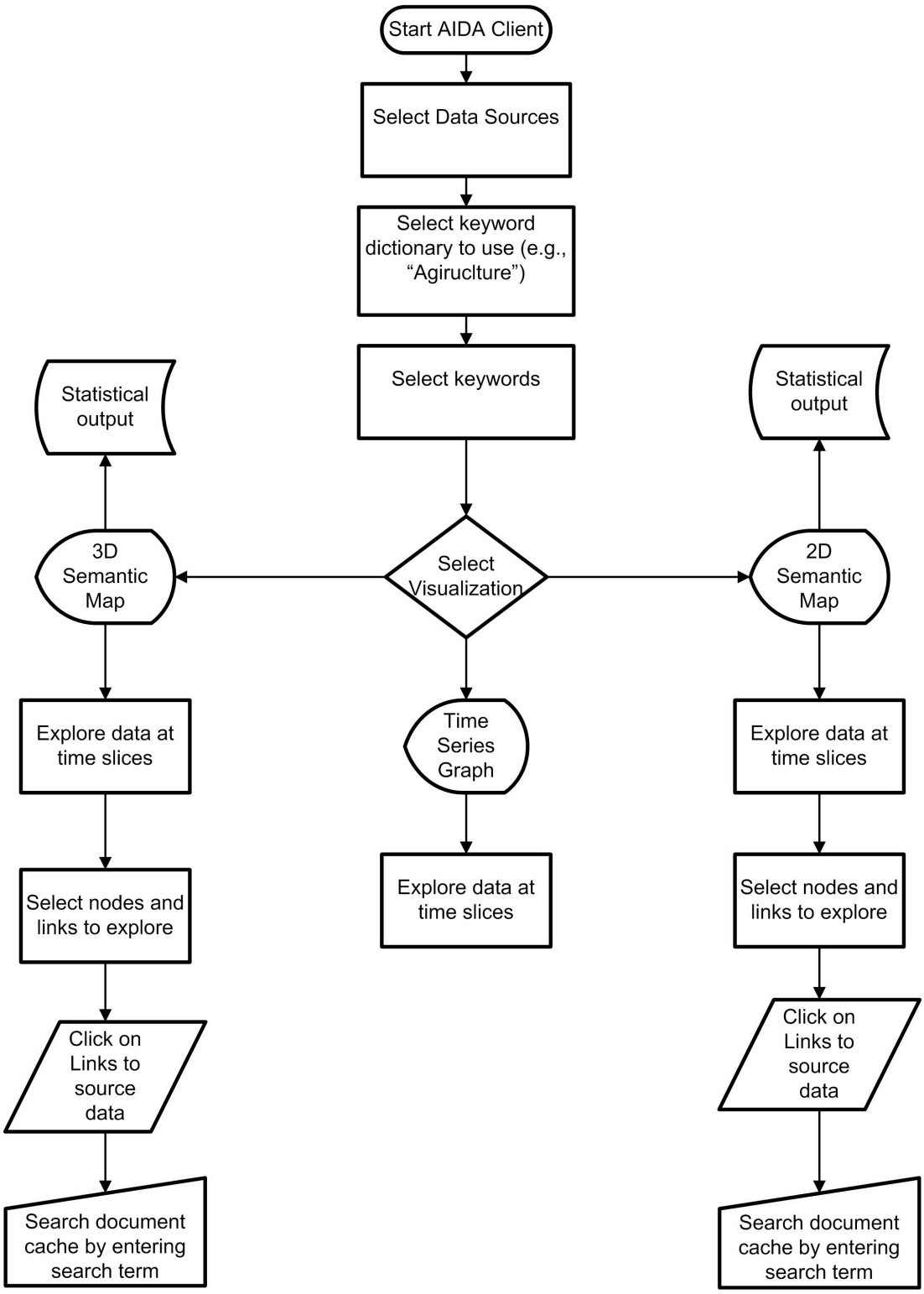

Figure 1. AIDA Client workflow

AIDA Client Layout

The AIDA Client can be started within any Java IDE. The code has been provided both for the client and server components so that users can see what functionality currently exists and the full implementation. See the "Getting_Started.pdf" file included with this download to see how to launch the AIDA Client.

Function 1: Select Data Sources: After starting the AIDA Client, the AIDA welcome screen opens and two main parts, a panel on the left side for structured data (including options for structured data analysis) and a blank main panel on the right open. This main panel is left blank because current data have only addressed structured sources. On the bottom left side of the screen, the Data Sources tab can be selected to switch the screen to the unstructured data sets, which are the focus of this manual. This will open the screen shown below (Figure 2).

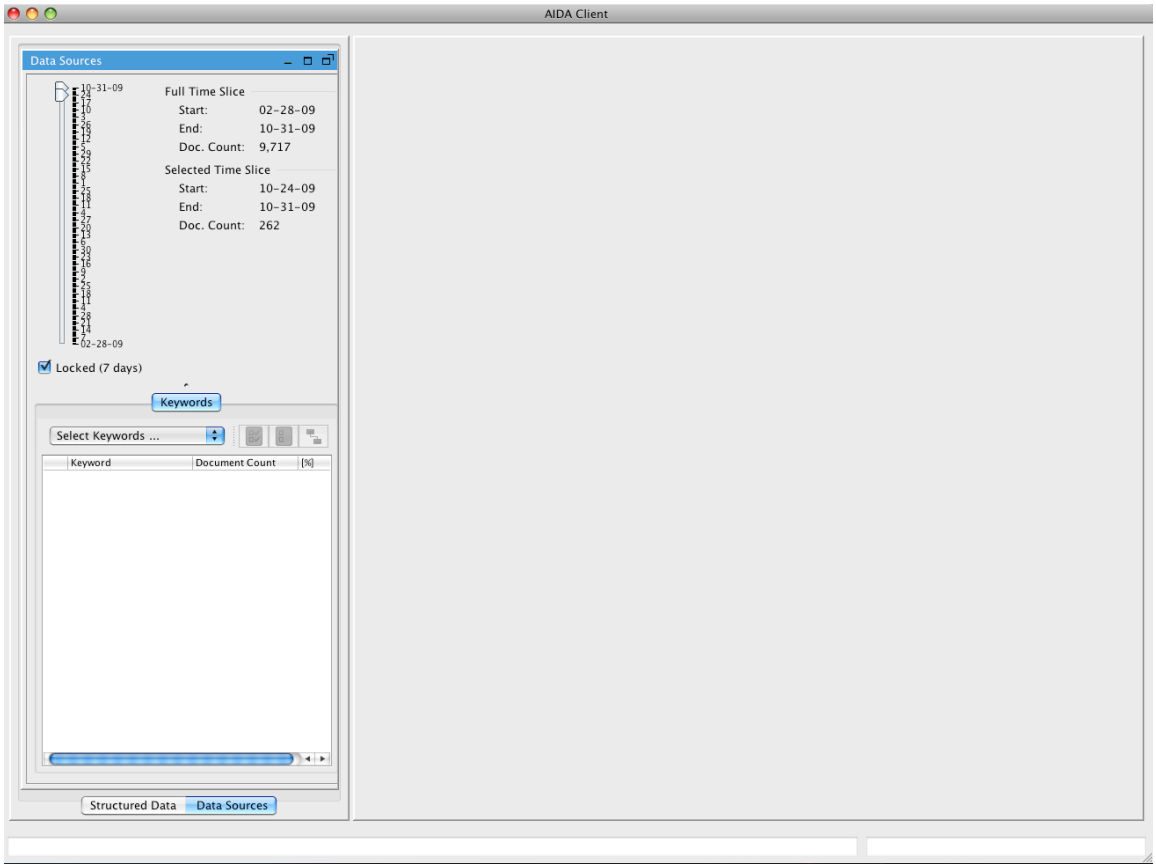

Figure 2. AIDA Client's Data Sources screen.

Function 2: Select Keyword Dictionary: Analysts choose data for visualization by selecting from the bottom left Keywords panel (in Figure 2) the term dictionaries (or categories), which are noticeable after pressing on the "Select Keywords…" chooser (Figure 3). The keywords selection process begins with the user deciding which of these lists he or she would like to explore. Note that the "Dynamic" category does not currently have any keywords and thus is excluded from discussion.

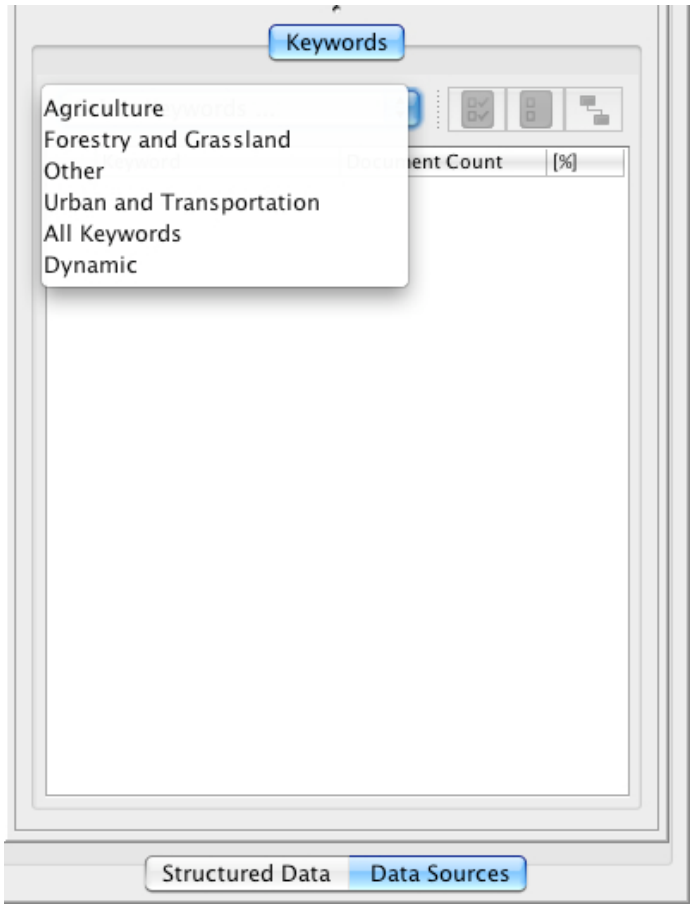

Figure 3. Keyword dictionary selection.

Function 3: Select Keywords: When the user has selected a list (e.g., "Agriculture"), a table containing all of the keywords in the selected list appears in the space below the selection box. By default, all of the keywords are selected when the list appears. Users have the option of narrowing their search further based on their interests. For example, an analyst who works on issues related to agriculture may only be interested in terms within the "Agriculture" category. Refining the search terms can be done through the following:

1) Uncheck the boxes to the left of those keywords to be excluded individually,

2) Press the uncheck boxes option (middle button on top-right button panel in Figure 4) and then select the desired terms.

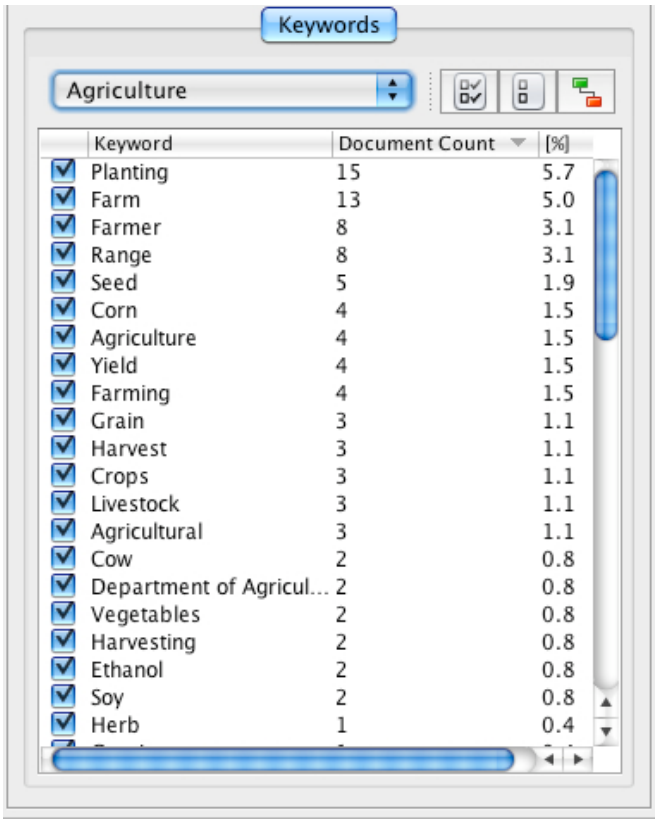

Figure 4. Keyword list

Function 4: Select Visualization: Once all keyword selection criteria have been defined, the user can then choose the visualization button. It has a picture of two linked boxes, one green and one red (Figure 5).

Figure 5. Visualization button.

The Configure Visualization menu will then appear (Figure 6). Defining a title for the visualization is optional. If the user plans on having more than one visualization open at a time, it is helpful to give each a name related to the data that is to be displayed (e.g., “Agriculture” vs. “Urban and Transportation” searches) or topic searched.

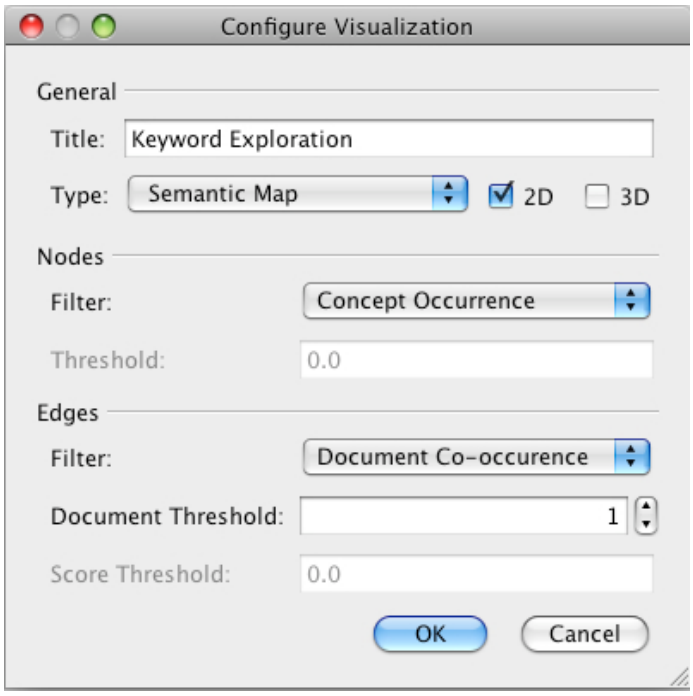

Figure 6. The Configuration Visualization menu.

There are three options for displaying the data visualization for selected keywords: Time Series, 2D Semantic Map, or 3D Semantic Map. First, the user selects the visualization type using the "Type" drop-down menu.

Time Series data can only be displayed in 2D. However, the Semantic Map visualization is available in both 2D and 3D. The user may also adjust the sensitivity of the network generator used in the semantic map visualization. This is accomplished through the filter type for individual nodes and the filter type and document threshold for edges (links). The network generator operates most efficiently when based on the "Concept Occurrence" option; other options include tf-idf, semantic similarity, and term frequency count. Once the "Type" has been selected, the user can then click the OK button in the Configure Visualization window. This opens the visualization selected based on the selected keywords.

Time Series: Time Series data visualization (Figure 7) is with a line graph of document counts. The x-axis represents time slices from oldest to most recent moving from the xy intercept. The y-axis is the number of documents in which a selected keyword is used. The max, min, and current counts of a keyword are provided in the keyword key below the graph.

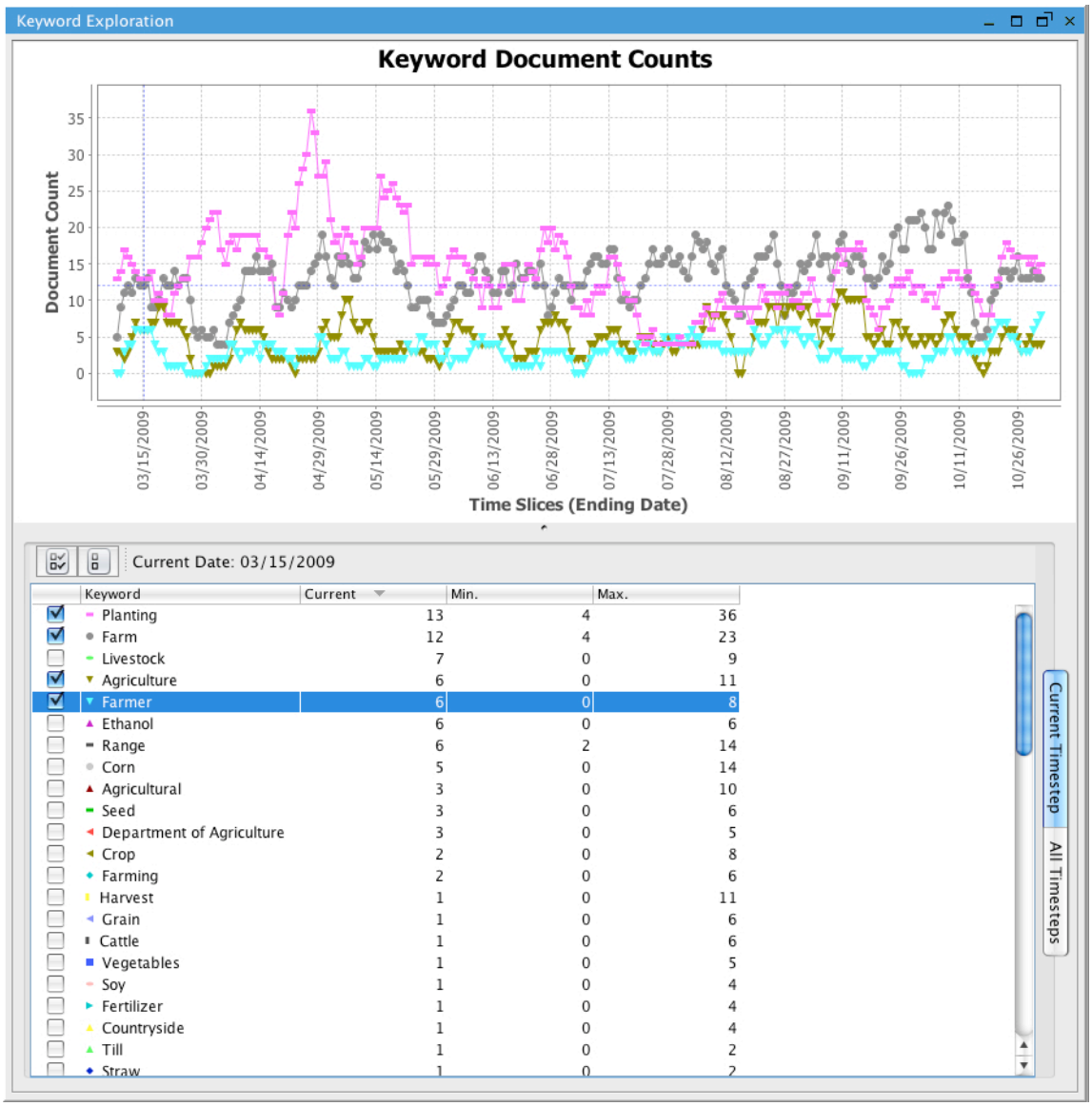

Figure 7. Time Series visualization showing keywords (bottom) and line graph (top) during different time intervals (x-axis).

Hovering over a point on the graph with the mouse arrow will bring up a comment displaying the document count, keyword, and time slice information for that point (Figure 8).

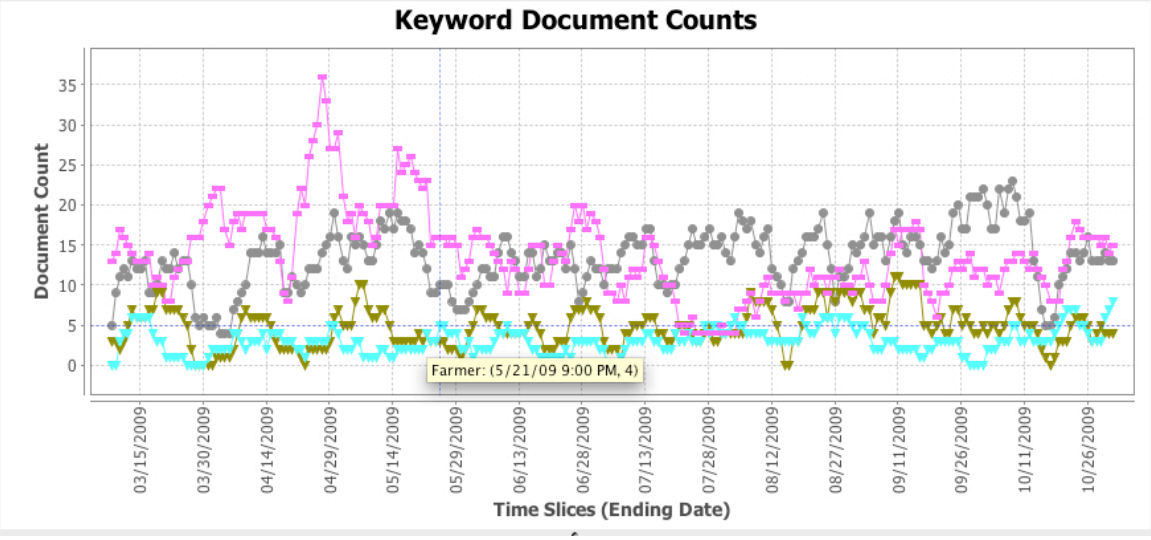

Figure 8. Time Series data details.

Below the graph is the legend (Figure 9). The Time Series visualization opens the default values in the legend at the oldest time slice. Clicking on the graph will select the time slice closest to the point selected and the legend will display the values for that time slice. The user can sort the legend according to the values in any of the columns. The user is also able to refine the search terms displayed on the graph using the legend. If only four terms based on the Max. Count are wanted, the user can sort the legend by clicking on the header for the Max. Count column, sorting from highest to lowest. The user can check or uncheck boxes on the left side of the legend at any point in the assessment.

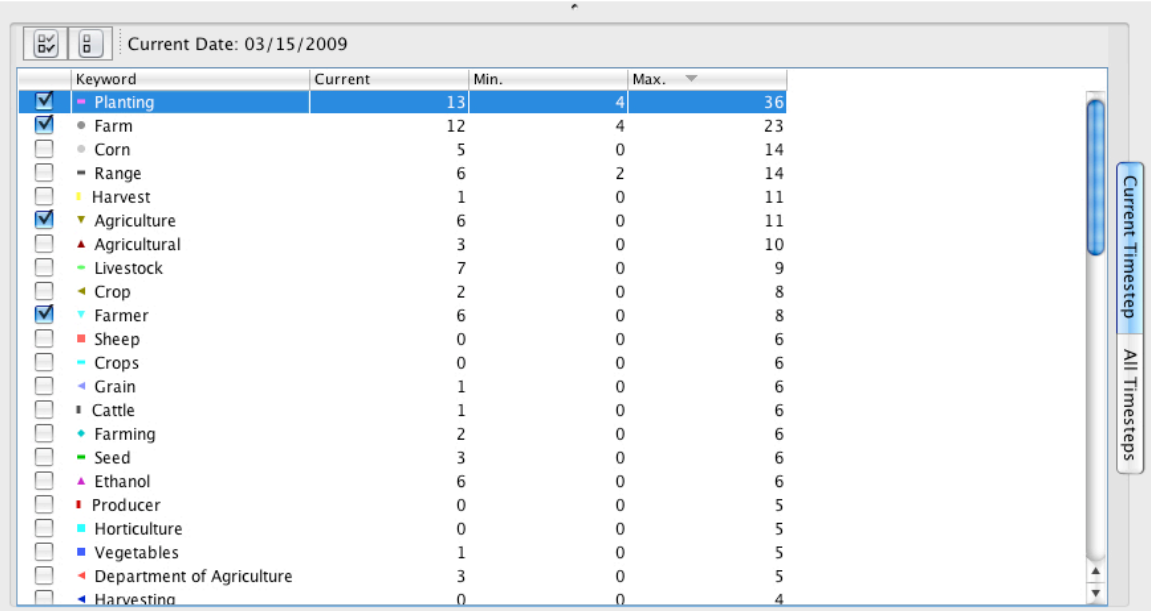

Figure 9. Time Series graph legend.

Semantic Maps

As stated, semantic mapping is a method of visually organizing the relationships between pieces of information. Keywords are displayed as nodes in the graph space using a yellow disk icon. The disks vary in size based on the number of keyword occurrences in a time slice relative to the mean number of times the term occurs in the cache up to that point in time. Lines connect terms in instances of their co-occurrence within or between documents. Relative strength of term links during a search interval is expressed by the following:

$$v_{oi} = \frac{l_{oi} - \min(l_o \in L_o)}{\max(l_o \in L_o) - \min(l_o \in L_o)}$$

with v being the strength of a link (i) at time interval o, l representing the number of linked documents, and L is the set of all links. This basic algorithm allows links to be valued at specific intervals against all other links, showing which terms have stronger or weaker associations with other terms.

2D Semantic Map: The 2D semantic map displays a single time slice of information as a semantic graph with nodes and links (Figure 10). When the map is opened, the most recent time slice is displayed.

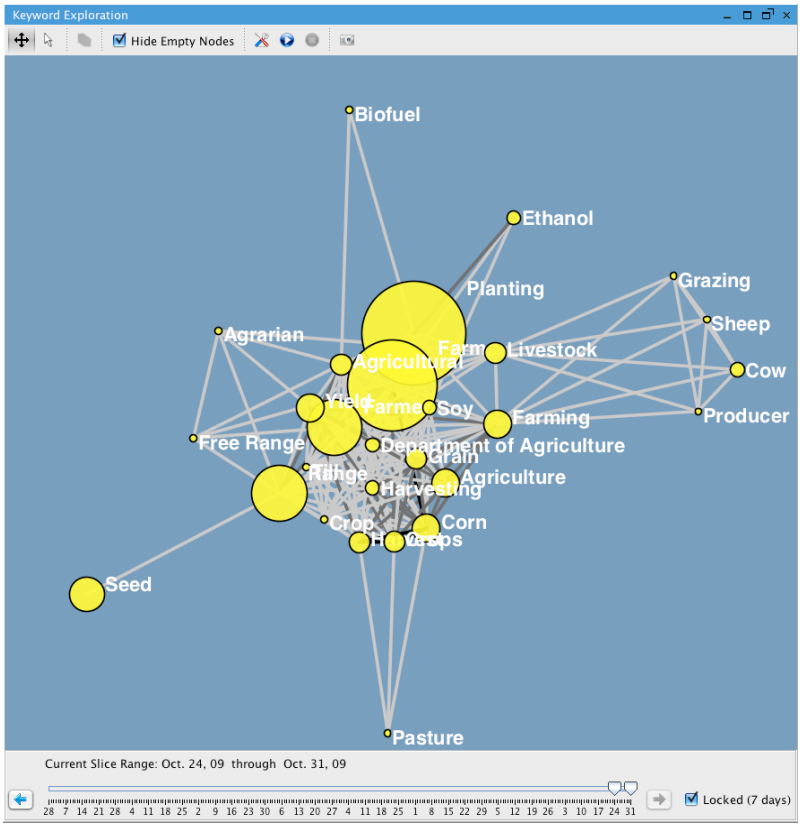

Figure 10. 2D Semantic Map visualization.

Although the visualization is only a single time slice, the entire cache of information is available for analysis. In order to be able to view all of the data, the user must run the

network generator. By clicking on the blue play button (top selection bar in Figure 10), the user starts the network generator, which queries the cache for the selected keywords and displays them, showing the nodes and links one time interval at a time. The user can move between time slices by using a slider selection tool at the bottom of the graph space window (see the bottom slider in Figure 10). Arrow buttons at either end of the slide bar move the time interval selector.

Left-clicking and holding on a node, or the link between two nodes, will bring up a floating window that displays basic statistics related to the term at that point in time: the date of the time slice, the document count, and the edge count (Figure 11). The document count is the number of documents in that time slice related to the selected term and the edge count is the number of other keywords to which the selected keyword is connected in that time slice.

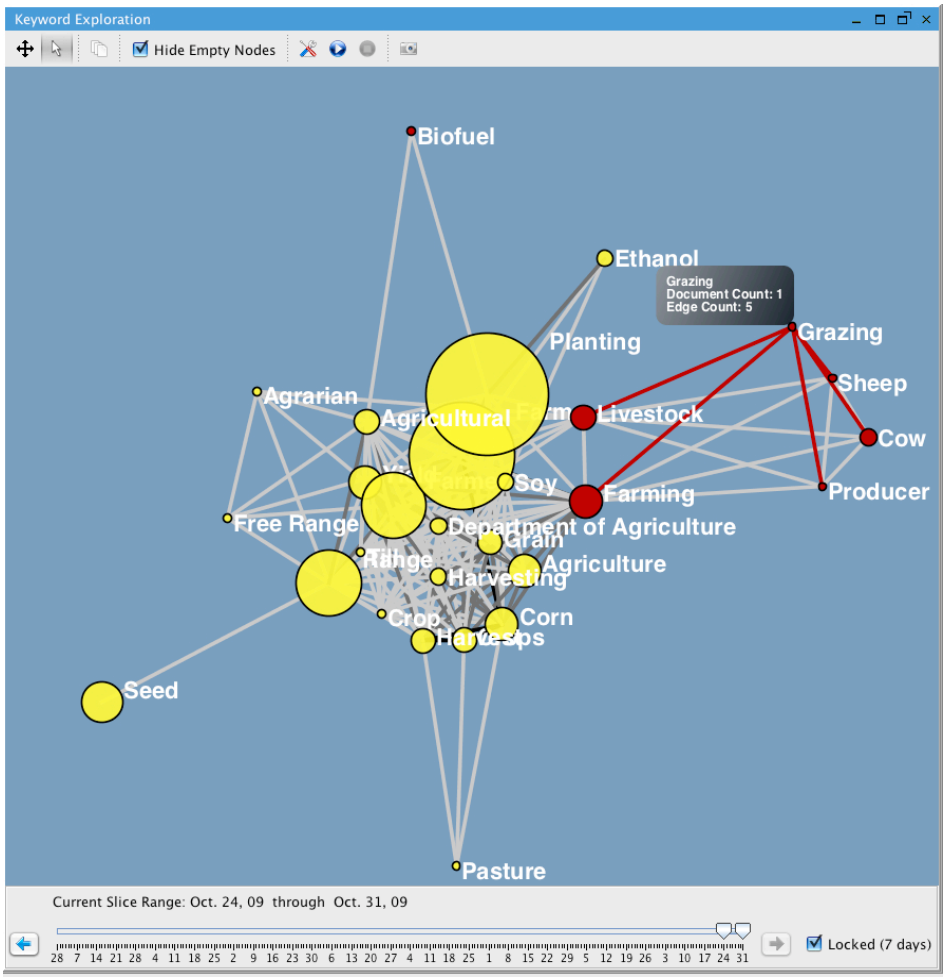

Figure 11. 2D Semantic Map node detail.

A single left-click on a node engages the document viewer function (third icon from the left in Figure 12), the icon for which is grayed out when no selection is made (see also the top selection bar in Figure 10). Selecting that icon will open the Document window. In the Document window, the headlines for documents related to the selected term at the

current time slice are listed (Figure 13).

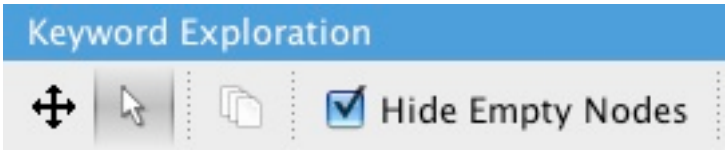

Figure 12. 2D Semantic Map document viewer control and selection bar.

Figure 13. Document window.

3D Semantic Map: The initial view in the 3D semantic map visualization is the same as in the 2D option. By clicking on the blue play button, the user starts the network generator, which queries the cache for the selected keywords and displays them and their associated links. The main difference in this visualization than the 2D map is that the

time intervals are stacked one on top of the other (Figure 14).

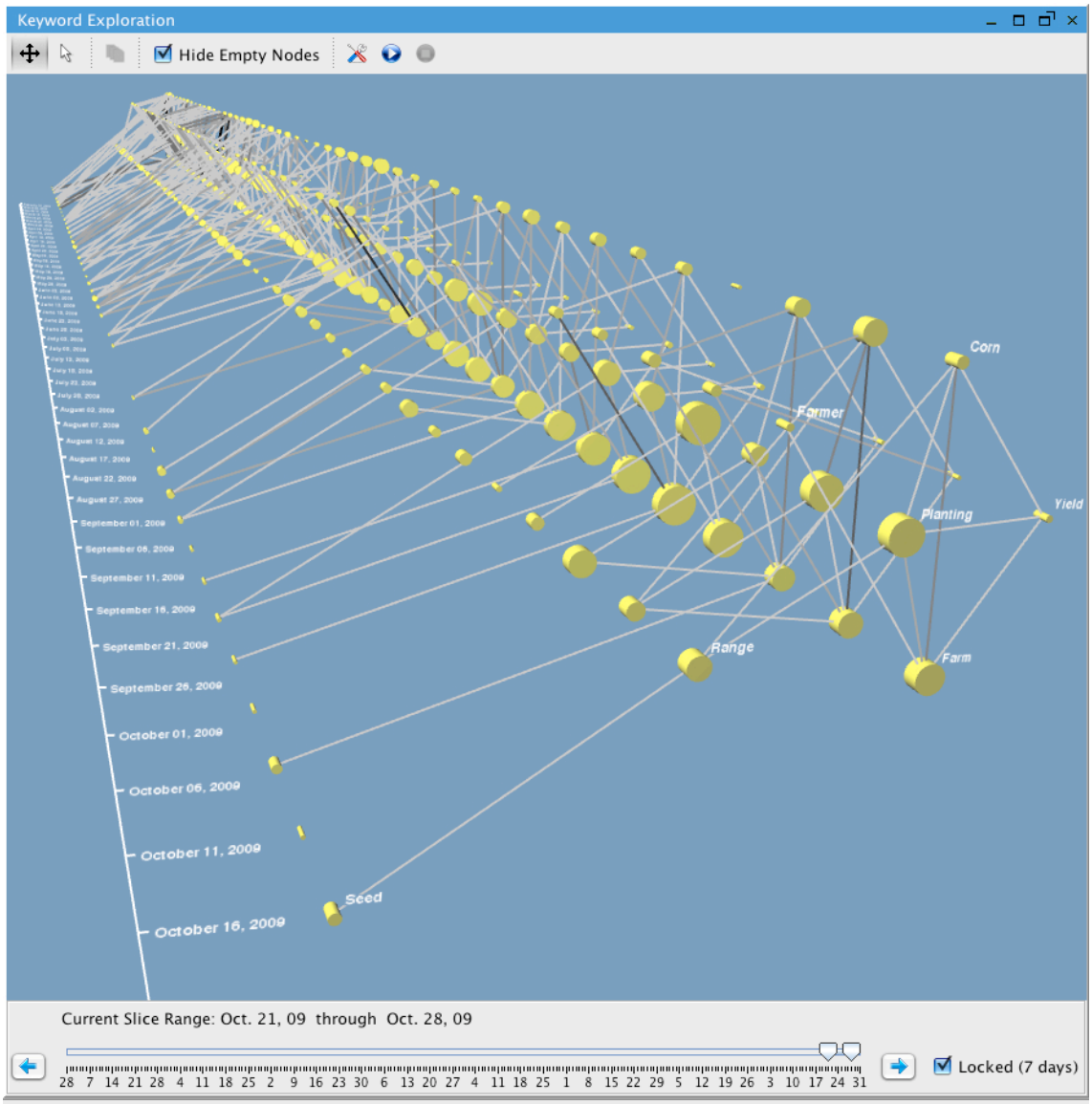

Figure 14. 3D Semantic Map. Note that only a few terms are displayed in order to making viewing easier.

When the network generator finishes running, the user's vantage point with respect to the data is looking straight down the middle of the stacked disks. In order to get a better perspective, the graph space can be tilted. This is accomplished by clicking on the move icon (Figure 15; see also Figure 12 far left icon) and then placing the curser anywhere in the graph space by left-clicking and holding. By moving the mouse from side-to-side or up and down, this will pivot the time axis around the most recent time slice.

Figure 15. Move icon.

Right-click and holding will slide the entire graph space from side-to-side or up and down. When finished with adjusting the perspective, the user should click on the selector arrow/curser button (Figure 16; see also second icon from left in Figure 12). This will allow the user to select nodes, links, and the Document window.

Figure 16. Selector arrow/cursor button.

At the bottom left of the graph space is a scale showing the time slice for each set of nodes (see graph in Figure 14). The user can move between time slices by using a slider selection tool (see bottom slider in Figure 14). Just as in the 2D Semantic Map, left-clicking and holding on a disk, or the link between two disks, will bring up a floating window that displays the date of the time slice, the document count, and the edge count (Figure 17). Again, a single left-click on a node engages the document viewer function, the icon for which is grayed out when no selection is made. Selecting that icon will open the Document window. The Document window lists the headlines for documents related to the selected term for the current time slice.

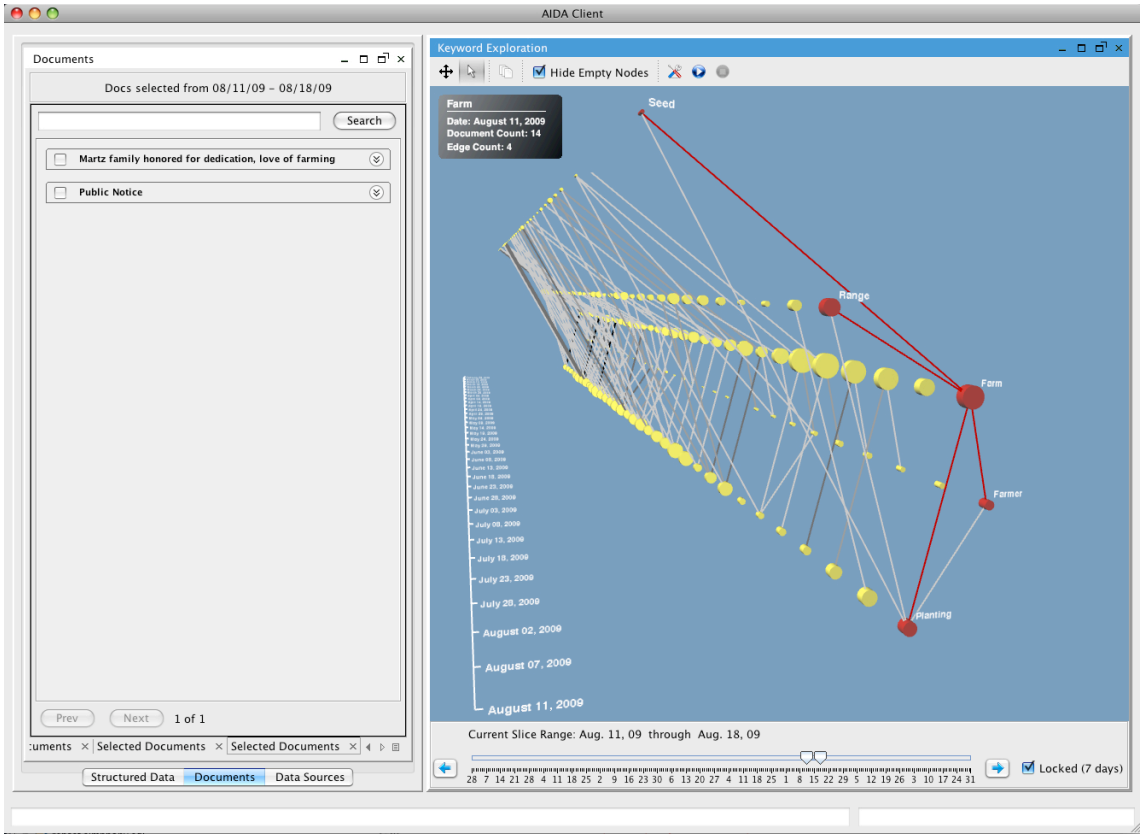

Figure 17. 3D Semantic Map Document window and selected nodes. Note that only a few terms are selected for visualization.

Statistical Output: For both 2D and 3D semantic map selection, the client will produce text files (CSV delimited) that contain counts for node terms, links, and documents per time slice. As discussed earlier, these files are outputted to the /Client/text_cache_output folder. Only terms selected for viewing will be outputted to the folder. Statistical tools (e.g., R) can be used to conduct further analyses on the output.

Function 5: Explore Data at Time Slices: For all visualization tools discussed earlier, users should click on the different time slices offered by the 2D and 3D options. This allows viewers to see which terms have greater frequency rates and links to other terms. This initial exploration could be used to provide the total assessment needed to understand specific land use trends; however, more refined analysis maybe needed.

Function 6: Select Nodes and Links to Explore: Once potentially significant trends have been identified, more refined searches and term relationships should be conducted (e.g., see Figure 17). This is simply done by only checking the relevant terms in keyword lists (see Figure 4) and running the visualization as needed.

Function 7: Click on Links to Source Data: As stated earlier, the Document list window is opened by using the mouse to interact with nodes and links in the Semantic Maps visualization window. Clicking on a node or link in the semantic map activates the document list icon at the top of the visualization window (Figure 18). After selecting a node or link in one of the semantic map views, the user may further explore the documents in which the keyword(s) represented by the selected node or link are specifically used (Figure 19).

Figure 18. Document viewer button.

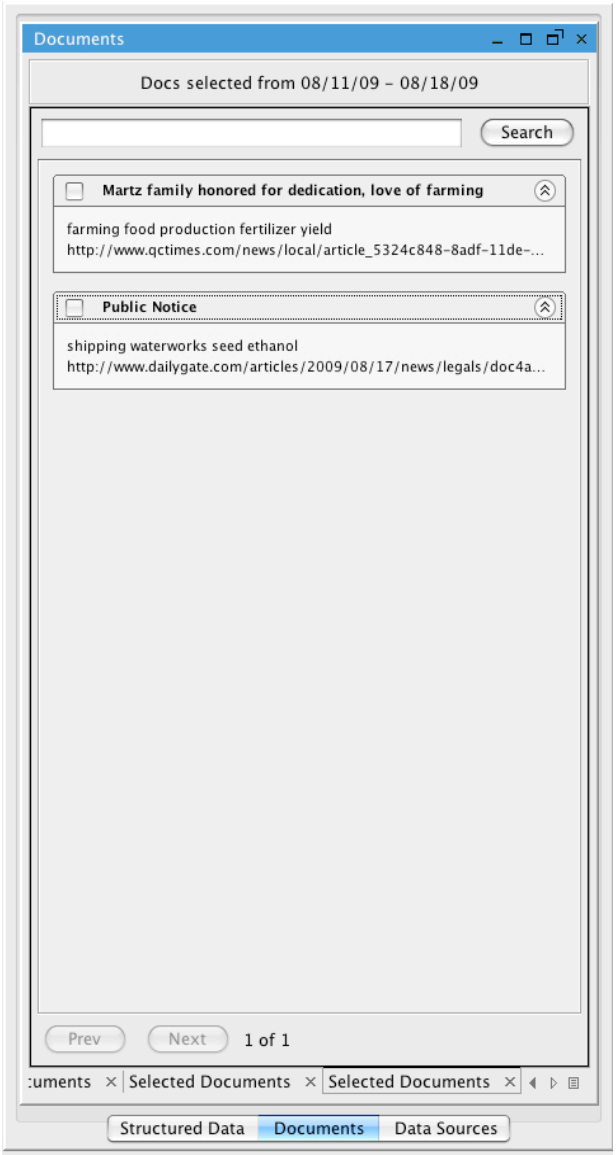

Figure 19. Document window based on the link between *farm* and *planting*.

The document exploration window opens as a collapsed list of headlines. Clicking on the chevron to the right of the headline will expand that individual box to reveal a very short summary of the document and the web link, if available (Figure 19).

Clicking on the web link will open the users' default web browser and navigate directly to the website page where the document is located (Figure 20).

Figure 20. Web browser opened by clinking on the top link in Figure 19's Document window.

Function 8: Search Document Cache: The document search function is located at the top of the Document window (Figure 21). A curser appears when the user clicks in the search box.

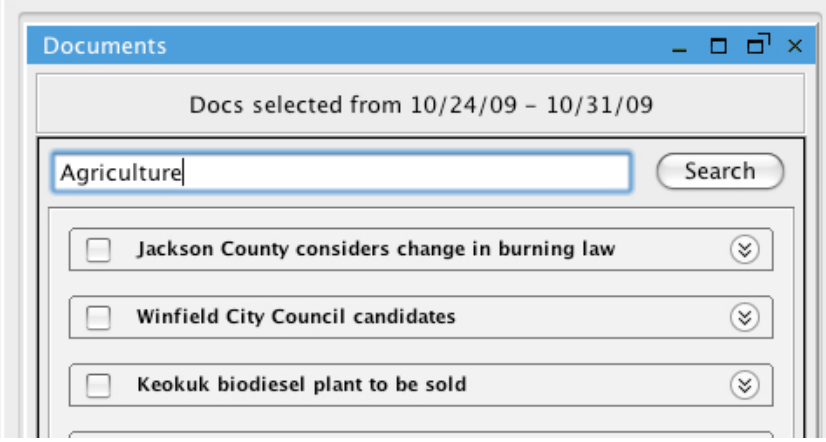

Figure 21. Document window search function.

The user can then type in search term(s) of their choice and click the search button. Any documents included in the current Document list are searched. Results are displayed in a new tab in the Document window, showing only the subset of the searched documents in which the term(s) appear(s) (Figure 22). If there are no matches a pop-up window is displayed:

Your search '[term(s)]' did not match any documents.

Unless it is told otherwise, the search is performed on all documents in the active Document list. Check boxes next to headlines in the Document list can be used to select a subset of documents in the list to be searched. The list generated through a search works just as in the document exploration view. Headlines can be expanded to see a brief summary, which includes an active web link to the source.

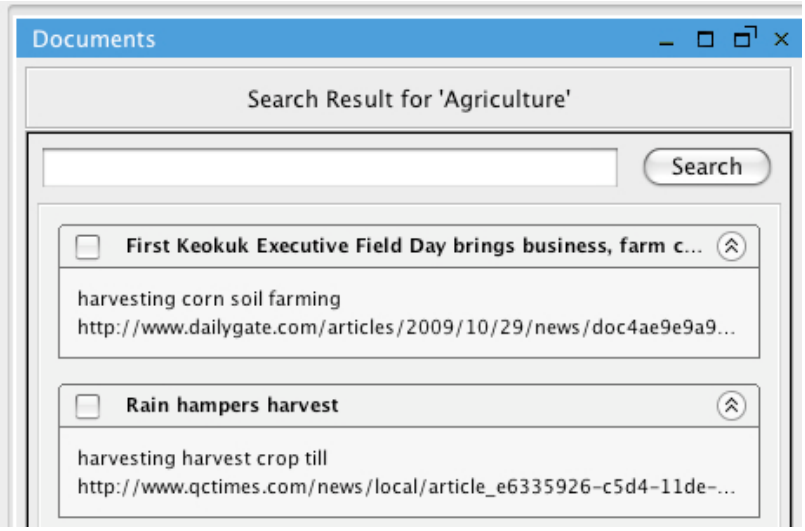

Figure 22. Document window returning search results.