Sustainability **2010**, *2*, 1182-1203; doi:10.3390/su2051182

OPEN ACCESS
sustainability
ISSN 2071-1050
www.mdpi.com/journal/sustainability

Article

Monitoring Land Use: Capturing Change through an Information Fusion Approach †

Mark R. Altaweel [1,*], Lilian N. Alessa [2], Andrew D. Kliskey [2] and Christopher E. Bone [2]

[1] Computation Institute, University of Chicago, 5640 S. Ellis Avenue, RI 405, Chicago, IL 60637, USA

[2] Resilience and Adaptive Management Group, University of Alaska Anchorage, 3101 Science Circle, Anchorage, AK 99508, USA; E-Mails: afla@uaa.alaska.edu (L.N.A.); afadk@uaa.alaska.edu (A.D.K.); afceb@uaa.alaska.edu (C.E.B.)

† All authors contributed equally to this work.

* Author to whom correspondence should be addressed; E-Mail: mraltawe@uchicago.edu; Tel.: +1-001-773-702-1895; Fax: +1-001-773-834-3700.

Received: 10 March 2010; in revised form: 6 April 2010 / Accepted: 27 April 2010 / Published: 29 April 2010

Abstract: Social and environmental factors affecting land use change are among the most significant drivers transforming the planet. Such change has been and continues to be monitored through the use of satellite imagery, aerial photography, and technical reports. While these monitoring tools are useful in observing the empirical results of land use change and issues of sustainability, the data they provide are often not useful in capturing the fundamental policies, social drivers, and unseen factors that shape how landscapes are transformed. In addition, some monitoring approaches can be prohibitively expensive and too slow in providing useful data at a timescale in which data are needed. This paper argues that techniques using information fusion and conducting assessments of continuous data feeds can be beneficial for monitoring primary social and ecological mechanisms affecting how geographic settings are changed over different time scales. We present a computational approach that couples open source tools in order to conduct an analysis of text data, helping to determine relevant events and trends. To demonstrate the approach, we discuss a case study that integrates varied newspapers from two Midwest states in the United States, Iowa and Nebraska, showing how potentially significant issues and events can be captured. Although the approach we present is useful for monitoring current web-based data streams, we argue that such a method should ultimately be integrated

closely with less managed systems and modeling techniques to enhance not only land use monitoring but also to better forecast and understand landscape change.

Keywords: monitoring; land use; information fusion; social-ecological; data mining; modeling

1. Introduction

As rapid and slow anthropogenic and ecological changes continue to affect many landscapes and vital natural resources around the world [1], the ability to create adaptive strategies to address the adverse effects of such change is becoming more critical. Prior to determining what adaptive strategy can best address this change in any given setting, relevant social and ecological drivers need to be identified. Environment and ecosystem managers, researchers, scientists, and other stakeholders are often tasked with identifying singular and reoccurring events affecting land use issues [2,3]. Remote sensing and other physical data are often used in monitoring these issues, but such data often do not provide information on social mechanisms or policies that are potentially relevant. This diminishes the ability for stakeholders to identify important behaviors or factors. Other approaches, such as technical reports, are often slow and expensive to conduct, leading to untimely information. Techniques, therefore, need to be developed that allow individuals to monitor significant events and event patterns. Such monitoring should identify both social and environmental drivers that affect land use change.

This paper introduces an information fusion approach that searches textual sources in order to identify singular events or longer-term event patterns. We focus on events that are caused by social and ecological conditions that are either directly or indirectly relevant to land use. We define indirect influences as those that relate to policy, legislation, decisions, social behavior, and other, often unseen mechanisms that can affect how land is used over time. The goal of this paper is to demonstrate that the presented information fusion approach is useful for identifying singular events and event patterns, which might be difficult to identify using other methods that potentially shape pertinent land use transformations.

We present our approach using a case study from the Midwest region of the United States, deriving information from online newspapers. Rather than being a primary focus of this paper, this case study simply serves as an example of how information can be captured and potentially useful terms and events assessed. Ultimately, we propose that our approach be applied to less managed systems and integrated with modeling approaches in order to best understand land use dynamics. We also provide access to the software developed in the hope that it can be further investigated by the reader and used for other efforts similar to what is proposed here. We begin the presentation by providing an overview of data mining and information fusion, describing how they can or have been used to understand events related to land use. We then present how the information fusion technique and tool we apply can be used for searches. Our case study is then presented, showing how events and factors affecting significant issues in the Midwest of the United States can be studied and monitored using the discussed approach. In the discussion, we detail the significance and future plans of our approach, providing

some general suggestions on how it could be used in less managed settings and integrated with modeling techniques in order to improve forecasting that attempts to better understand land use dynamics.

2. Research Methodology

2.1. Overview of Data Mining and Text Analysis

Data mining and text analysis have been used for a variety of fields, including those relevant for social-ecological systems (SES) and land use change [4-6]. Methodological advancements include the use of artificial intelligence, statistical procedures, and algorithms for monitoring the significance and rates of terms [7]. Other techniques include those that compare [8] source texts to reference text corpora in order to identify emerging term patterns that enable one to better comprehend human perceptions of noteworthy news stories. Some researchers have applied techniques that use text mining to assess the risks of certain events or even detecting anomalous patterns of events [9,10]. With these types of methodological advancements, event monitoring and decision support tools become more feasible for analysts and other stakeholders. More tools are beginning to appear that can have potentially significant benefit for researchers in detecting event behaviors and patterns over different time scales [11,12], with these tools potentially applicable for detecting events affecting land use.

Despite these achievements in data mining, relatively few tools and techniques have been developed that allow significant land use events to be monitored at both the local and larger regional scales using a variety of social-based signals that can incorporate web-based data streams. In addition, the tools and techniques developed often lack capacity to integrate a variety of terms and information sources in order to determine semantic patterns that may divulge important information on relevant events. Network-based techniques, however, can potentially provide a structure that enables analysts to monitor events not only based on specific terms but also events that are best understood through relationships between different terms. Text searches and information retrieval may need to provide data on patterns of different events and their relevance over relatively lengthy periods [13]. Added to this, single events, perhaps sometimes significant, maybe detected and determined by investigating relevant terms associated together; such terms may generally not be associated together in data sources but their sudden co-occurrence could potentially aid in the detection of important events. For instance, in a rural county in Illinois, a company recently attempted to begin the process of building wind turbines in an area that has significant wind resources [14]. However, many people in the local region are opposed to this decision because of the effect this will have on the surrounding region, specifically how it may change the physical appearance of the region and affect property values. Furthermore, if the wind turbines were built, the land use characteristics of the surrounding region may change, as farm property could be sold and new industry is attracted to the region. What the example shows is that a single event, *i.e.*, the decision to build wind turbines, may only be reported in an isolated or relatively few data sources, but multiple social- and environment-related words (e.g., *wind turbine*, *farm*, *land use*, *project developer*, *windy*) that are present in data sources can aid in the detection of a potentially significant event.

Whether analyzing long-term event patterns or specific events, a variety of social-ecological behaviors and decisions may clearly affect how landscapes transform and the potential resilience of

specific regions to change [15]. There is a need, therefore, for approaches and tools to be able to provide assessments of long-term and individual events that may affect land use practice. Searches that can utilize relationships between various terms, therefore, can be useful in detecting expected or unexpected events.

2.2. Overview of Information Fusion

An approach within data mining that integrates and couples multiple terms and data sources is information fusion. More specifically, this technique applies data from a variety of structured and unstructured sources, with specific linkages made between shared terms in data sources [16,17]. A common usage of information fusion is in conducting searches that are focused on specific topics of interests but have a variety of relevant terms and data sources. Frequencies of terms and displays of semantic linkages are outputs that can be produced by the approach [18,19]. Analyses using information fusion can vary, and different semantic assessments could be appropriate depending on the type of search. The benefit of information fusion is that multiple terms, or potentially different types of media data, can be displayed together for their relationships. As an example, a text search may have a primary term (e.g., *agriculture*) but important secondary subjects (e.g., *crops*, *urbanization*, *legislation*) may form links with the primary term search. This creates the possibility to display trends by visualizing primary and secondary terms, including their co-occurrence and relationships over a given period. Analytical and qualitative visual approaches can be used to indicate rarely linked terms, which may delineate significant stories despite rare links, and more common term linkages. The example of the wind turbine article demonstrates an event that was not widely reported but potentially has relative significance for the particular region, as economic structures and land use affecting farmland in the surrounding region could be dramatically altered by the introduction of wind turbines to the region. Term searches on the primary term (e.g., *land use*) along with secondary terms (e.g., *wind turbine*, *project developer*, *farm*, *windy*) help focus the search to ultimately find this event.

Although tools such as Google Trends [20] provide graphical perspectives of term frequencies and event statistics, semantic relationships between terms are not easily displayed. This can be limiting in finding events of relevance from retrieved data. In our approach, we utilize and display collected data by creating term relationships in a semantic network. Stakeholders potentially require the identification of singular events as well as aggregated information over longer periods to display important land use trends. Our approach allows both of these by using common statistical procedures and identifying links between terms using visual display at different temporal scales. This type of approach allows the analysis to incorporate both quantitative and qualitative assessments.

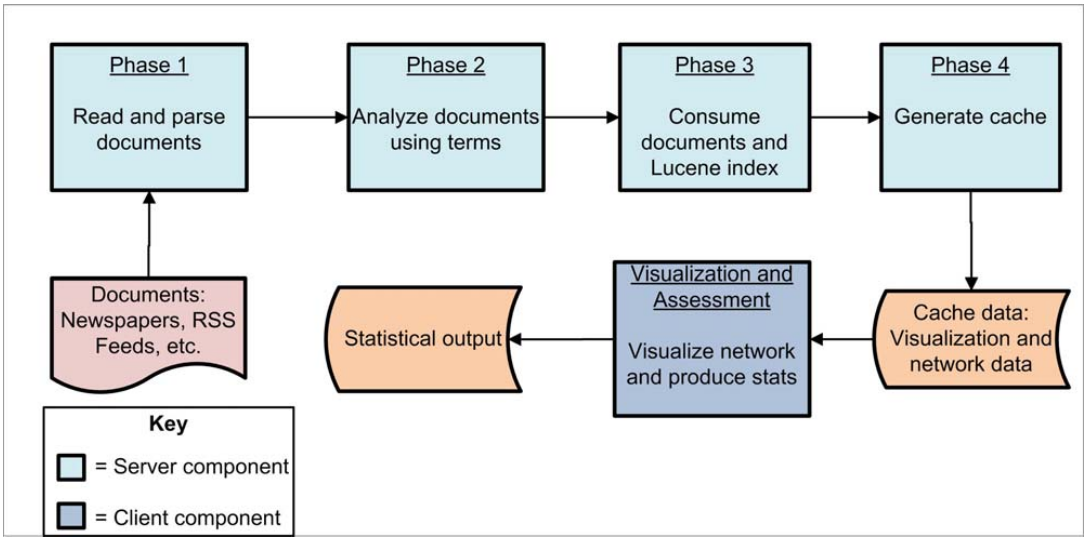
3. Applied Methods

This paper demonstrates how a network-based perspective that links terms and term frequency over time is useful in identifying pertinent events as they are observed. To facilitate this demonstration, we present information on our applied tool, which is called Architecture for Integrated and Dynamic Data Analysis (AIDA). Some of the methods and concepts in AIDA have been described elsewhere [5,21]; we will summarize these as well as describe additional new approaches currently applied. Readers can access this tool by downloading the code; manuals for using the server and client components are

provided, with the case study and associated search data as an example application included in the download [22]. The documentation included has more detailed explanation on how AIDA conducts term searches, analysis, and user-defined output.

The description below on our applied method provides a step-by-step explanation of how relevant terms are found and semantic relationships formed. Examples for each step are provided in order to facilitate the discussion. The method has four primary phases that retrieve, organize, and produce data that are searchable for identifying events. These four faces are applied within the server application of AIDA. After these four phases, we then apply the visualization and assessment client that displays the results and produces statistical text output that could be used for further quantitative analysis. AIDA only uses open source tools in all of the applied phases. In addition to investigating AIDA's documentation, readers should also assess the third-party tools referenced in the description below for further details and documentation on those tools. Figure 1 provides an overview of the four phases applied in the server's search procedure that produces an output for the client component of AIDA.

Figure 1. The client and server components and analysis in AIDA.



3.1. Phase 1

The first phase applies the UIMA [23] collection reader in order to read, parse, and reference data found in online sources. UIMA facilitates the analysis of unstructured content such as texts. This process includes the tagging of necessary metadata from an article (e.g., *Golden Triangle Newspapers* article). Documents can be read in a range of formats, including ASCII, HTML, SGML, XML, RTF, Email, PDF, and Word format. The metadata include a long integer UTC timestamp in milliseconds, the document URL, the title, and the author(s) of the article if available. These delimited data are then used for the article analysis and tagging conducted in Phase 2.

3.2. Phase 2

The second phase applies the General Architecture for Text Engineering (GATE) that analyzes and tags relevant terms in documents (e.g., *wind turbine, farm*) that are defined by the AIDA user [24]. In addition to terms defined by the user, AIDA integrates WordNet, which finds term synonyms called synsets that can be searched along with the user's selected terms [25]. The user can edit the output from WordNet in order to accept or reject searched synonyms. To summarize the processes in this phase, text searches are conducted by analyzing and using Language Resources (LR), Processing Resources (PR), Visual Resources (VR), and Natural Language Processing (NLP). In addition to these tools, the architecture applies CREOLE (Collection of REusable Objects for Language Engineering). To summarize, all of these tools enable un-annotated document content to be passed and parsed, with terms (*i.e.*, those provided by the user) identified and relevant annotation tagging conducted. Users in the phase provide a list or database of terms and their synsets found by WordNet.

The second phase applies a sequential set of PRs that conduct text annotation procedures (Figure 2). The first of these PRs is the AnnotationDeletePR, which removes previous annotations from an LR as it is read. In addition to this PR, the DefaultTokenizer PR splits text into simple text tokens (e.g., words, number, and punctuation). A third PR applied is the SentenceSplitter PR, which splits text into sentences, applying a gazetteer of abbreviations that distinguishes end-of-sentence punctuations from other punctuation marks. The next PR is the DefaultGazetter PR, which is used for annotating text through plain text lists of terms. Terms in lists are distinguished by different lines and a gazetteer feature separator that enables specifications for text annotation. An example of the format of this specification is:

sunflower%cName = Sunflower

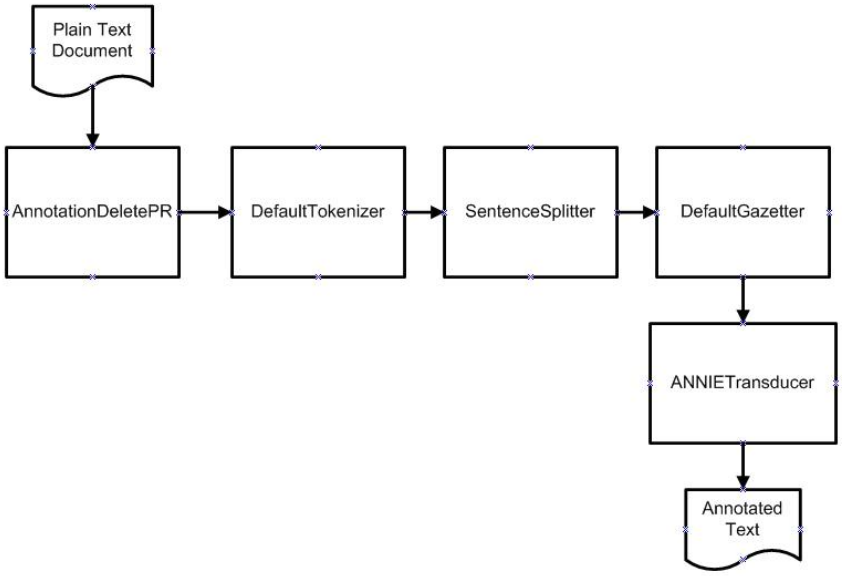
helianthus%cName = Sunflower

In this example the identified common terms (*sunflower* and *helianthus*) use a feature separator character ("%") and apply a GATE feature name ("cName") for searches on the canonical term *Sunflower*. Capitalization is not necessary in this case, with both capitalized and un-capitalized forms annotated to allow searches to aggregate different terms based on the fact that they reference a common meaning. The final type of PR used is the ANNIETranslator PR, which applies the Java Annotation Patterns Engine (JAPE) rules for annotation. Specifically, JAPE allows words to be associated within a larger semantic context that associates like words. For example, text can be annotated to be organized using typologies (e.g., a category defined as "Agriculture") with terms (e.g., *sunflower*) placed within the category. In this example:

```
Rule: Agriculture
(
{Lookup.majorType == agriculture}
)
:agriculture -->
:agriculture.Agriculture = {kind = "Agriculture", canonicalName =
:agriculture.Lookup.cName, rule = "Agriculture"}
```

defines the category annotation with the "kind" feature set as "Agriculture" and the "cannonicalName" equivalent to the cName feature (e.g., *Sunflower*) of the lookup annotation mentioned earlier. This example allows terms such as *Sunflower* to be associated with a larger set of terms that are encompassed within "Agriculture." By placing terms within larger categories, it is possible to search specific topics based on sets of terms relevant to land use categories such as "Agriculture". This allows search results to aggregate term searches by categories or differentiates searches within categories. Word content is also lemmatized, which groups different term inflections to be analyzed as a single item, in searches using TreeTagger software included in GATE [26].

Figure 2. The sequential application of the PRs for enabling text annotation.



Terms chosen to be searched may derive from reference texts, which could be domain expert databases, term lists constructed from user-specific knowledge, or other sources. Regardless of the source, derived reference texts should comprehensively address the subject domain (e.g., agriculture). Stakeholders in land use, as an example, should provide relevant terms (e.g., *agriculture, farming*) that are commonly used by the data sources searched. In the example to be discussed, common terms that relate to land use and are used by newspapers are listed in the data provided for download. Relevant terms can be determined by a preliminary investigation of common terms used by articles through an initial search of terms and then expanding the initial list as other terms are found; the use of WordNet aides in this endeavor.

3.3. Phase 3

Phase three is called the consumption phase, whereby terms found in documents are indexed using a Lucene index [27,28]. Lucene is a text search engine library; its primary use here is for providing an index to store data (e.g., term frequency) on terms searched. Supplementary to the explanation provided here, readers should investigate the description provided by the open source Lucene tool and

documentation. Using Lucene, the common and canonical annotations extracted from Phase 2 are indexed with their metadata information derived in Phase 1; this includes the timestamp, URL, title, and any known author(s). Document contents (*i.e.*, terms searched) are referenced in the index; the location information provides the user with the ability to click and access specific documents from the final cache phase (*i.e.*, Phase 4) during the visualization and assessment step. Whole documents are not stored in the index, as this would needlessly require additional memory and storage space in the visual analysis. Instead, frequencies and term counts are stored.

The resulting index then allows the scoring of the relative importance of terms in documents. Currently, terms are scored for their relative value using tf-idf [29] scoring that applies the following formula:

$$S_{td} = TF_{td} * IDF_t * N_d \tag{1}$$

where S is the tf-idf score of a term (t) in a document (d), TF is the term's frequency, IDF is the inverse document frequency, and N enables a normalization factor or boost value of d to affect S regardless of d's length. In the case study, all N values have the default entry of 1.0. In Equation (1), TF is evaluated by:

$$TF_{td} = \frac{n_t}{\sum_i w_i} \tag{2}$$

with the number of instances (n) of t divided by the sum total of all term (w) occurrences in a document. The final coefficient used for the tf-idf score, IDF, is defined as:

$$IDF_t = 1 + \log\left(\frac{|D|}{|\{d : t \in d\} + 1|}\right) \tag{3}$$

where D is the total number of documents and d represents each document containing the term (t). Not only can index scores (*i.e.*, tf-idf) be used to identify the relevance of a term in a single document, but the IDF variable can be applied independently for assessing the significance of a term in all available documents. The IDF value provides a higher score for more rare terms.

3.4. Phase 4

The final phase creates a cache of all the documents found in the search. The cache is created using the Lucene index from Phase 3. The cache is relevant for a Google-like search of located documents in AIDA's visualization tool (*i.e.*, see Section 3.5). The search is enabled using a collection of ScoredTermCollection (SCT) files, which are a type of data storage that facilitate searches of the cache and hold term data. In the cache, terms can be organized using distinguished categories; this allows different term counts to be added together and included within the same category designation if desired. Categories, in essence, act as term dictionaries that contain the list of keywords for land use topics investigated. XML is used to specify the organization of the SCT. In addition to term data contents, the SCT files incorporate the duration of the cache and time interval information. The search data include the start date, the total number of days searched, and the range of each search interval.

3.5. Visualization and Assessment

STC files enable index scores (*i.e.*, tf-idf) and other basic statistical information to be stored and used in determining the frequency of searched terms. In essence, STC files become the database for the search. In visualizing and assessing the cache's data, terms can be displayed in a network whereby terms are nodes that are linked with other terms that co-occur at a specific time interval. The importance of a node can be displayed using its 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. Relative strength of term links during a search interval is expressed by the following:

$$v_{oi} = \frac{l_{oi} - \min(l_o \in L_o)}{\max(l_o \in L_o) - \min(l_o \in L_o)} \quad (4)$$

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. This potentially helps to filter relevant and non-relevant terms in searches by providing output of terms that only have links with other terms. In other words, terms that have multiple links with other searched terms are more likely to be relevant to the goals of the search (e.g., stories dealing with land use change). In the figures showing the case study's semantic networks, weak v is shown as light colors and darker colors represent stronger links. Counts and frequencies are displayed within nodes and links for quick visual reference. In addition to visual displays, statistical information, including term counts, the number of term links, and number of documents per time interval are provided in text output (CSV format) in order for the data to be used in analysis. This output is used in the quantitative analysis of the case study described below. Additional analytical capabilities include GIS techniques, using GeoTools [30], which can be applied if structured data are available for assessing the relationships of term data to locations.

4. Applied Case Study

Changes to landscapes caused by anthropogenic and environmental influences have led to unexpected consequences affecting communities and natural resources at both rapid and slow time scales [31]. Because such changes can be of significant importance, relevant events and event patterns need to be identified by managers, researchers, and other stakeholders. The following case study will demonstrate our developed tool through the use of local online newspapers that potentially capture events affecting land use issues.

In the case study, we apply Phases 1–4 twice. We initially begin by producing a list of relevant HTML sources to search for a given time period. The end result of all of the phases is a searchable cache used in the visualization client as well as relevant text output that can be used to assess term and term links in statistical packages such as R. The first search finds and builds searchable terms using a wide time range (2006–2009); the second search conducts the specific period analysis using the terms found from the first search. Although the initial search may miss some rarely occurring terms, many relevant secondary terms (e.g., *plowing*) or broader categories (e.g., *agriculture*) that are applicable for land use topics (e.g., search on the loss of farmland to urban expansion) can be found. WordNet is used

after the initial search is conducted, as it enhances relevant terms found by providing their synonyms. With the incorporation of different types of terms listed for categories and the investigation of term relationships, rare and recurrent events are more likely to be located in searches.

Our case study focuses on land use change issues affecting Iowa and Nebraska. After a preliminary search, we found that agriculture, urban development, transportation, forestry, and grassland issues are significant topics of interest. In addition, there are other issues, including those related to the weather, legislation, and energy, that directly or indirectly affect land use change. The initial search resulted in obtaining a list of 300 terms of potential relevance. Roughly one-third of the terms searched are terms we derive from articles based on the initial search, while the rest are terms found using WordNet. To focus this example after the preliminary search, we conduct a specific search on a 245-day span from March 1, 2009 to November 1, 2009. This search involves the following online newspapers: *Daily Gate City*, *Le Mars Daily Sentinel*, *Golden Triangle Newspapers*, the *Muscatine Journal*, *Quad-City Times*, *Sioux City Journal*, the *Columbus Telegram*, the *Lincoln Journal Star*, *Omaha Newsstand*, *York News-Times*, the *Fremont Tribune*, and the *Grand Island Independent*. The first six papers are based in Iowa; the last six are Nebraska newspapers. These papers are chosen because they cover different regions within each state. They cover not only the immediate vicinity of the towns they are based in but also cover regional issues that affect other parts of the state and even surrounding states. Because exact geographic information is difficult to obtain and the data are unstructured, specific geographic references to terms and events are difficult to derive. Based on this, we did not apply GIS techniques in the example provided. In addition, we only searched the local and regional news stories provided in the papers. Future work on search algorithms applied, however, may allow our approach to better incorporate unstructured data with GIS.

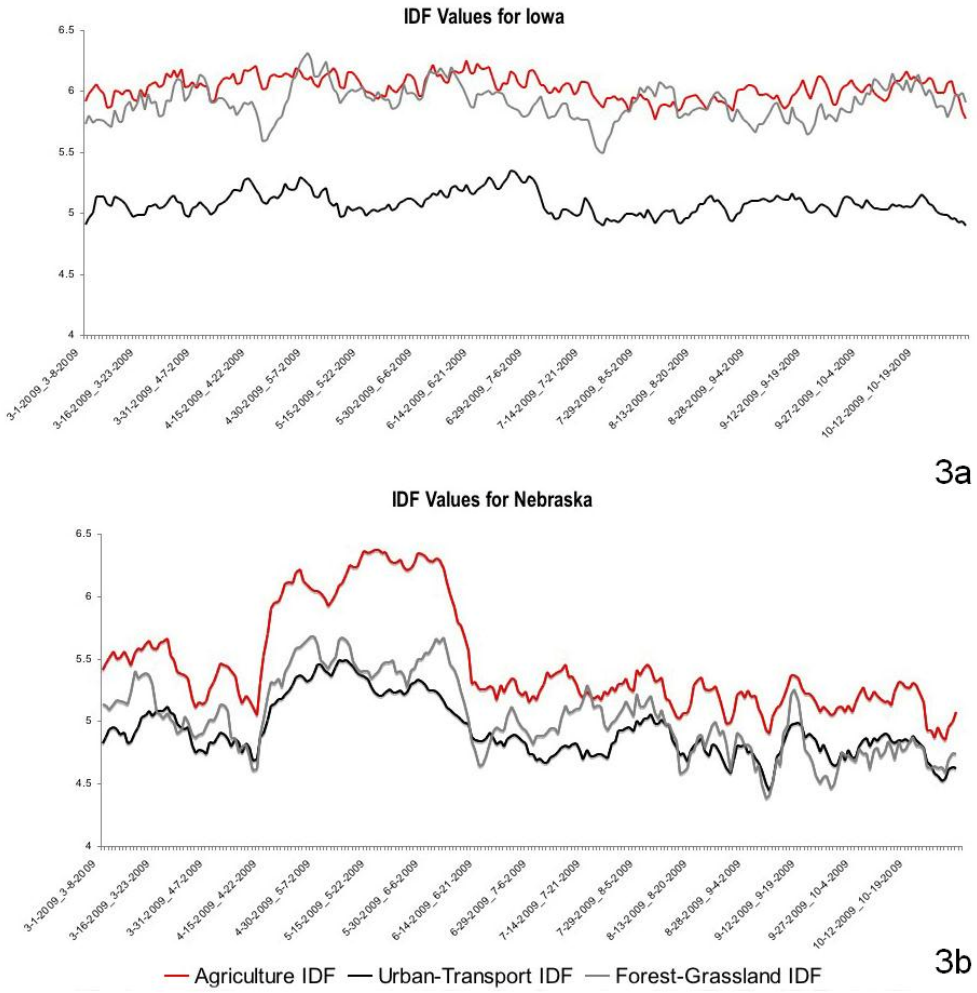
The total number of sources searched is 35,067 unique articles. For simplicity, the scenario is divided into two sub-scenarios, specifically dealing with Iowa and Nebraska. There are three primary categories created, which are designated as "urban and transportation", "agriculture", and "forestry and grassland." We also have a general "other" category, which focuses on environment, indirect (e.g., legislation) factors affecting land use, and other potentially significant land use issues (e.g., rural development). The terms in the "other" category could directly relate to the alternative categories mentioned, but because many of the terms in "other" are broad they are placed in this category. The AIDA download includes all of the terms used, with these terms organized within the four categories.

4.1. Case Study Results

Figure 3 provides the average *IDF* scores during the monitored period for the three main searched categories, which include the terms within these categories in each of the sub-scenarios (Figure 3a Iowa; Figure 3b Nebraska). For Iowa, terms related to urban and transportation issues are the most common throughout the search period; in the case of Nebraska, urban and transportation issues are common, but terms related to forestry and grassland topics are often just as common or are at times more common. We should note that there are far fewer terms related to the forestry and grassland category, which makes each term's *IDF* score more influential in the average *IDF* score for the category. In the case of Nebraska, one noticeable trend is that the number of documents retrieved spiked from approximately 100 per search interval to at times over 500 per interval from late April to

early June. This spike in the number of articles increased *IDF* (*i.e.*, leading to a lower rate of retrieved documents). We should note that the reporting of events does not always directly link to when the event actually occurred. The focus in the case study, therefore, is on when events are reported, as this indicates the identification of an event by a news source even though the event may have occurred at an earlier period.

Figure 3. Average *IDF* scores for the three main categories' terms in the Iowa (3a) and Nebraska (3b) sub-scenarios.



The relevance of terms and their relationships to other terms emerges more clearly when examining average *v* values in the different categories relative to the number of documents containing these terms (Table 1; Table 2). What can be noticed in these tables is that some less frequent terms have greater *v* values than terms that appear more commonly. This indicates that these greater *v* value terms are relatively highly linked with other searched terms. This linkage begins to indicate which terms potentially have greater relevance to the searched topics. The term *service*, for instance, is very

general, but it can have significance to certain searches related to land use issues. The results in Table 1 indicate that although *service* appears relatively frequently, it is not as commonly associated with other searched terms as less frequent terms such as *river*. Greater *v* values, in summary, allow the filtering of terms based on the strengths of term relationships. In other words, the more linked a term is with other searched terms the more useful the term becomes for land use categories. This automated approach does not fully remove non-relevant articles (*i.e.*, noise), but enables analysts to focus searches on terms that have the greatest linkage values (*i.e.*, the *v* values). From the topics in Tables 1 and 2, terms such as *community*, *street*, and *city* have the greatest rates of association with searched terms. This demonstrates the potential primacy that urban and transportation issues have in land use topics discussed by those local newspapers that were searched.

Table 1. Aggregate document count and average *v* values for terms in the Iowa sub-scenario.

Term	**Document count**	**Average v**
Urban and Transport		
Community	19,657	0.0386
Street	13,453	0.0429
Transportation	3,921	0.0166
Mayor	5,606	0.0221
Traffic	4,925	0.0222
City	31,982	0.0652
Agriculture		
Harvest	563	0.0066
Corn	1,322	0.0091
Agriculture	1,152	0.0072
Farm	3,173	0.0134
Planting	3,283	0.0132
Crops	403	0.0083
Farmer	722	0.0063
Seed	628	0.0045
Forestry and Grasslands		
Forest	447	0.0098
Wood	1,249	0.0161
Great Plains	16	0.0008
Prairie	1,219	0.0091
Other		
Service	13,387	0.0277
Council	10,260	0.0309
River	10,113	0.0315
Committee	19,657	0.0245

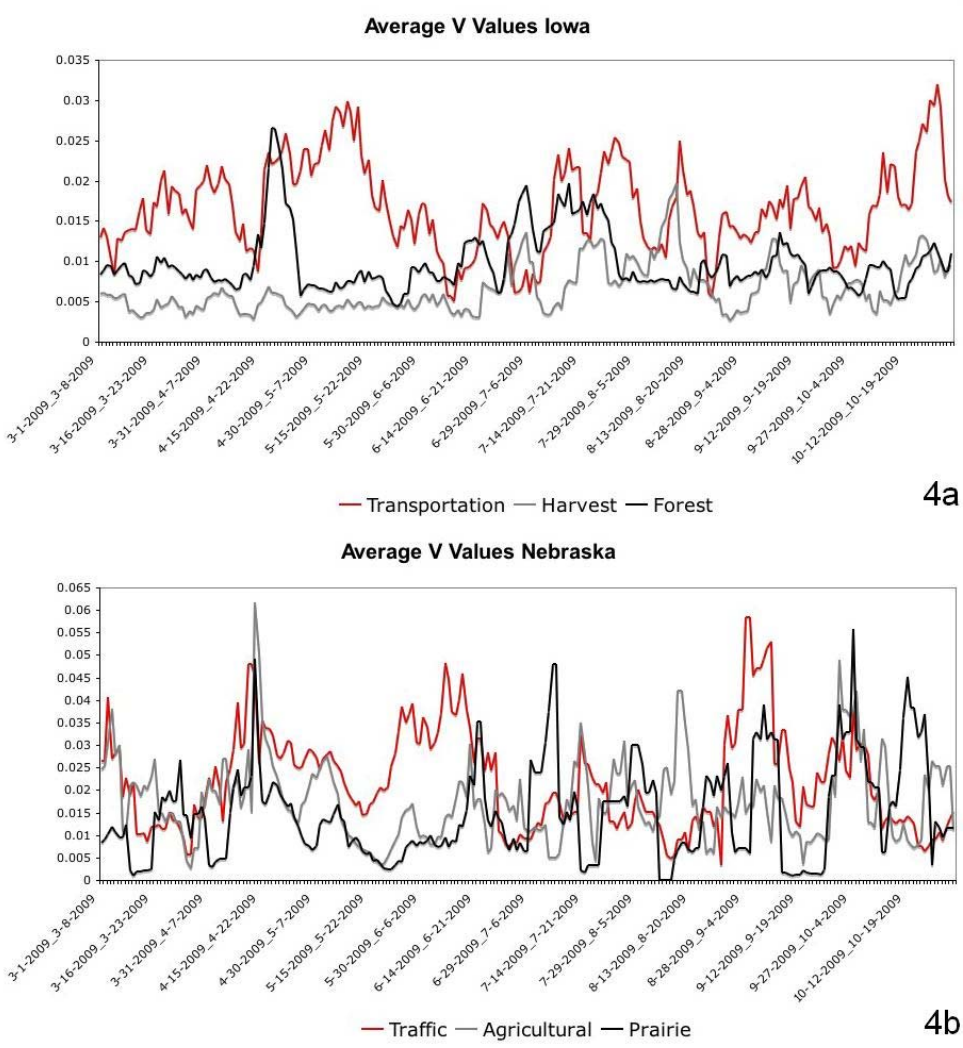
Table 2. Aggregate document count and average *v* values for terms in the Nebraska sub-scenario.

Term	Document count	Average v
Urban and Transport		
Community	9,664	0.0464
Street	6,956	0.0427
Highway	2,632	0.01844
Mayor	3,162	0.022
Traffic	2,175	0.0222
City	12,775	0.0638
Agriculture		
Grain	381	0.0125
Corn	955	0.0152
Agricultural	988	0.0044
Farm	2,069	0.0213
Planting	2,426	0.0162
Crops	441	0.0096
Farmer	523	0.011
Cattle	526	0.0119
Forestry and Grasslands		
Forest	296	0.001
Wood	861	0.0154
Plain	279	0.0063
Prairie	770	0.0148
Other		
Service	6,968	0.0373
Board	6,275	0.0336
River	2,103	0.0152
Bill	3,670	0.0288

By looking at specific time intervals within the overall search period, patterns of when different terms are reported in articles becomes apparent. Using the *v* value measure, we can see at what period terms become more linked with other searched terms. As an example, three terms (*transportation*, *harvest*, and *forest*) from Iowa-based newspapers are investigated (Figure 4a). During April and May, transportation issues become more relevant largely because state government increasingly debates transportation issues, federal stimulus funding for transportation projects becomes a major topic of discussion, communities increasingly debate and discuss planned road construction, and road construction activities and their reporting increase during the spring. After the initial increase in transportation stories, the issue subsides somewhat but then rebounds in its relevance to other searched terms during the late summer and fall. The majority of relevant stories concern road construction and closures. In June, the term *harvest* has greater prevalence with other terms as articles describe how weather conditions will affect the harvest. By August, an even greater number of stories mention how favorable summer weather has led to a good potential harvest in the fall. Although in September and

October *harvest* stories decline, the term still has a greater rate of linkages than it did in the spring, as several stories discuss weather conditions affecting the harvest. The term *forest*, in general, does not have a high rate of linkage with other terms; however, the linkage increases in April. During that month, Iowa's Department of Natural resources issues a report that, among other things, rates Iowa's ability in land and natural resource protection. The report discusses the state's role and policies in protecting its forests, with the report giving the state a B– in its land protection efforts and management. In July, *forest* again becomes somewhat prevalent, but the term never has as many linkages as it does during a brief reporting period in the spring.

Figure 4. Term *v* values during different periods for Iowa (4a) and Nebraska (4b).

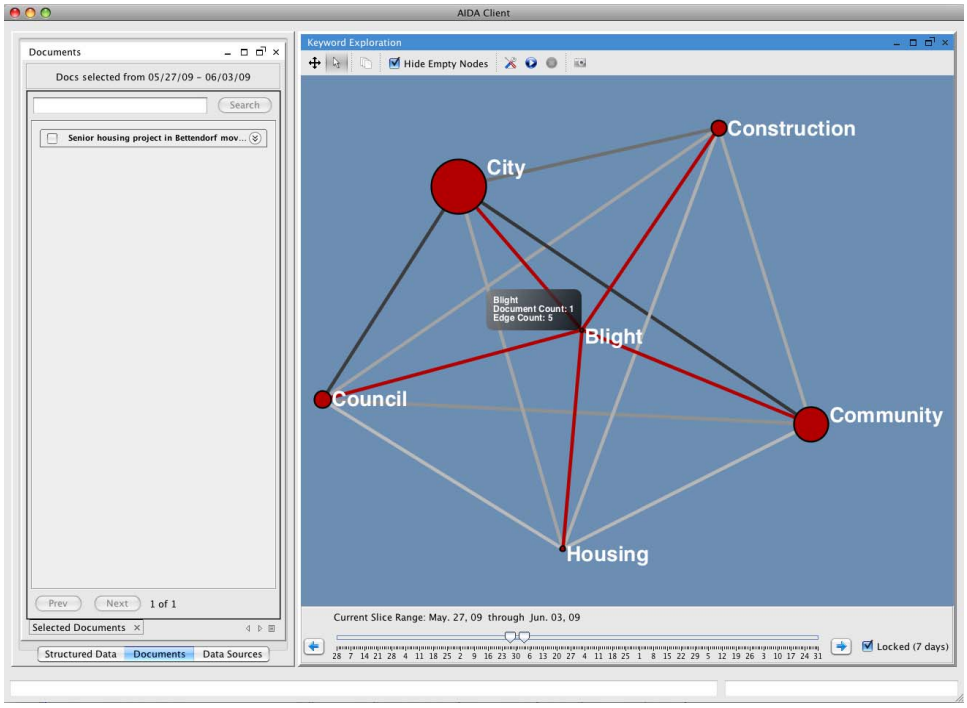


For Nebraska, the term *traffic* is similar to *transportation* in that the term has reoccurring relevance to other terms rather than being highly linked for only short periods (e.g., f*orest* in Figure 4a; Figure 4b). Similar to Iowa, many of the traffic stories throughout the searched period relate to road

construction. In April and May, a few of the stories discuss urban planning, neighborhood projects, and the construction of new factories in towns, which may positively or negatively affect traffic flow. In early September, one article discusses planned urban "greening" projects in Lincoln, Nebraska [32]. Such stories regarding planned or active urban projects along with road construction appear to cause *traffic*'s *v* to increase in different intervals during the search period. In April, stories containing *agricultural* produce a relatively high rate of linkages with other terms primarily because many of these stories relate to legislation or federal funding for projects that affect agricultural lands. These projects include water improvement and construction of sewage facilities for agricultural enterprises. Other stories include funding to exterminate infected livestock, compensation for lost land, and plans to construct livestock facilities. In August, articles discuss the cool weather affecting the coming harvest, increased federal purchases of pork, and the recession's impact on agricultural activity. These stories help the value of *v* for agriculture to increase during this period. In the early fall, stories containing *agricultural* often pertain to the imminent harvest. The term *prairie* has linkage rate peaks in April, July, September, and early October. In April, articles found include those that concern controlled burns of grasslands and prairie lands previously purchased for conservation. In July, articles reference a study that shows an increase in ponds in prairie lands and the subsequent increase in duck populations; other stories mention Nebraska's plans to change the state's Game and Parks Commission, potentially affecting conservation planning and ongoing efforts in prairies. In early October, articles discuss how a federal stimulus project affects prairie lands.

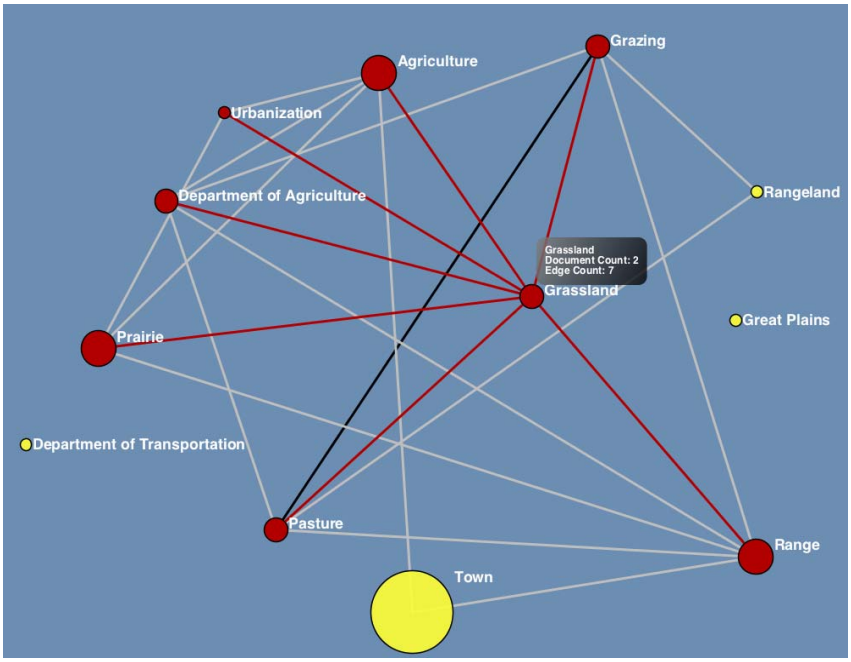
In addition to terms that may have increasing linkages at different periods, important events affecting land use may only be mentioned by a relatively small number of articles. As mentioned, network-based perspectives allow semantic relationships to be searched based on linkages with specific terms. When dealing with large sets of data that contain at least some of the terms searched, relationships between linked terms aid in determining which stories are of potential importance. As an example, events concerning urban renewal efforts in Iowa can be distinguished from other more common events based on the relationships of specific terms. In this case, the terms *community*, *city*, *housing*, *construction*, and *council* and their relationships with the term *blight* (Figure 5) help find a pertinent article on urban renewal. Automated searches can be focused so that results only return a group of terms (e.g., *housing*, *construction*) that co-occur with a primary term (*blight*), helping to provide identification of relevant events for narrowed searches. This example shows not only how narrowed searches can be found and visualized in AIDA, but clicking on the term links and returning the stories that contain linked terms can lead to further investigation of specific articles concerning urban renewal. In this case, the term *blight* has a tf-idf score of 0.0084 in the one relevant article found, showing the term's rarity in the article despite its pertinence to the urban renewal project [33]. Visual assessment and the use of tf-idf scores are also helpful in further refining a search once some filtering has been applied using the *v* values. In other words, because not all noise can be removed by *v*, visual assessment (e.g., Figure 5) and the significance of a term in a document may become necessary for refined searches.

Figure 5. Linkages of terms with *blight* help to distinguish an urban renewal event from other articles. The panel on the left returns the URL links to the specific stories found that have the linked terms or are single terms as distinguished by the nodes. In this case, only one article has the term *blight* and the five link connections indicated.



One major land use issue faced in Nebraska is the preservation of grasslands. Although earlier we mentioned that the term *prairie*, which is often related to *grassland*, has greater overall links in April, July, September, and October, other articles of similar significance occur in months in which *prairie* is less mentioned. Figure 6 shows a June 2009 example in which the term *grassland* has clear linkages with the terms *agriculture*, *grazing*, *range*, *pasture*, *prairie*, *Department of Agriculture*, and *urbanization*. The relationships between these terms assists in the location of articles that are less ambiguous with regard to their content and relevant for land use issues concerning grasslands. The articles that contain *grassland* and the terms that are linked to it pertain to federal conservation funding and the loss of habitat in Nebraska's grasslands [34,35]. These stories have tf-idf scores for *grassland* at 0.185 and 0.0323, which are significant results since *grassland* is often a rare term but has great relevance to the documents retrieved. More significantly, it is the connection of *grassland* with other terms that helps to narrow the search and return articles of significance for grassland issues.

Figure 6. Term linkages to *grassland* are evident for this Nebraska example.



In addition to 2D displays, 3D displays can assist in the location of not only rarely occurring and specific types of events based on term linkages but also aid in the ability to trace term patterns and relationships at different time intervals. For instance, one major story in 2009 is the use of federal stimulus money for a variety of state projects. In the case of Iowa, stimulus funding has been used for building construction and infrastructure projects. Even though in most years a term such as *stimulus* might be relatively rare, this term becomes more common in 2009 because of federal funding issued under the American Recovery and Reinvestment Act [36]. Tracking this and other related terms is not only possible by applying a 3D perspective, but the inclusion of specific terms and the filtering of others in a search enable specific topics to be found more easily (Figure 7). In other words, searches on the data cache can be narrowed to a limited number of terms in order to find specific topics related to the stimulus funding. As an example, in the spring many stories pertaining to the stimulus concern urban infrastructure projects. In later periods, articles concerning the stimulus address how funds enable road construction and facilitate other types of transportation infrastructure development. Figure 7 shows how the term *stimulus* has clear link relationships with *Department of Transportation*, *legislation*, *construction*, *transportation*, *repair*, and *road* during the July 8–15 interval. In the figure, earlier semantic relationships between *stimulus* and other terms are seen behind the highlighted *stimulus* term (*i.e.*, that from July 8–15). A similar tracking of this term can also be applied for Nebraska (Figure 8). In the case of Nebraska, during the June 5–12 interval, stimulus-related events pertain to village construction projects and water treatment/sewage projects that could be possible based on the funds. Terms relationships between *stimulus* and *road*, *committee*, *village*, *sewer*, *construction*, and *community* help find this set of events. From the perspective in Figure 8, *stimulus* is

seen to be increasing in usage (larger nodes at the back of the perspective) during the early spring in April, but by June and July the term is declining in use.

Figure 7. Term linkages with *stimulus* in this 3D example from Iowa.

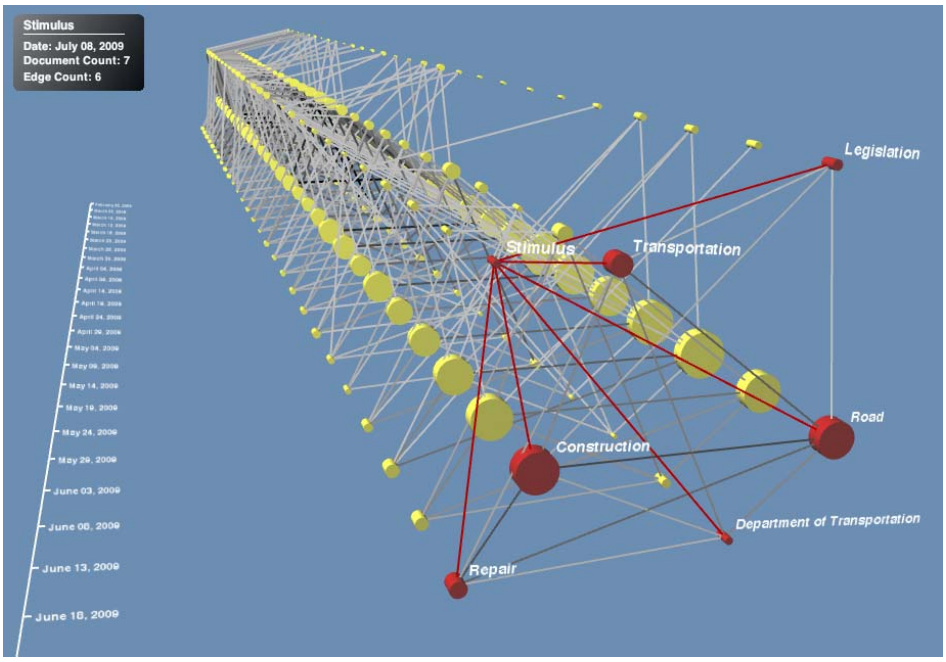
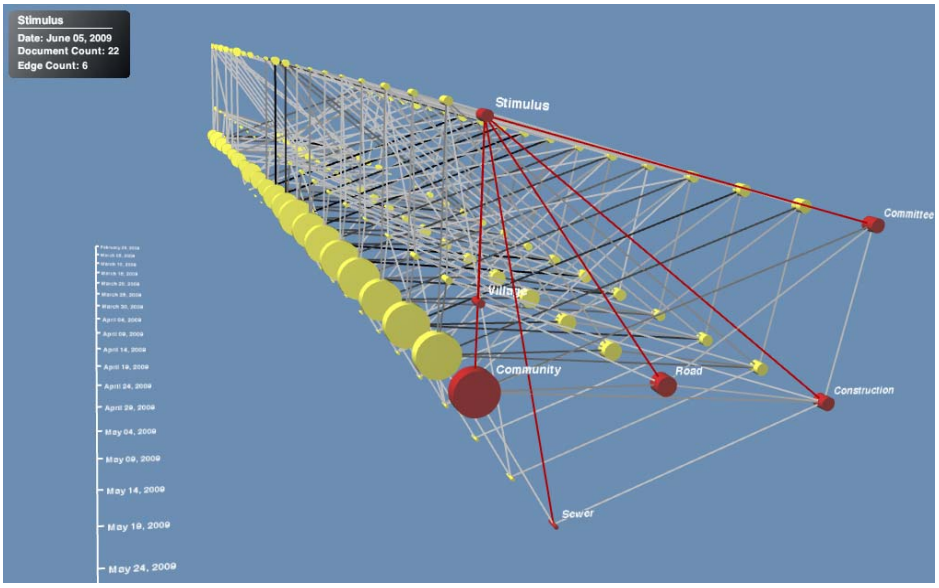


Figure 8. Term linkages with *stimulus* in this 3D example from Nebraska.



5. Discussion and Conclusions

Term searches that incorporate a network-based perspective of linked terms, term frequency over time, and visual and quantitative analyses are useful in identifying recurrent and rare social-ecological land use events. The ability to understand land use change and events affecting sustainability topics requires that both physical measurements, such as remote sensing and climate data, and social events that are not easily measured, including acts of legislation or policy decisions, to be concurrently monitored. In such cases, the approach presented provides analysts with the ability to conduct both quantitative and qualitative assessments that investigate patterns of relationships over time, determine the importance of terms and relationships between terms, and identify unique events that could be of significant interest to stakeholders. Analysts investigating events that can affect communities and natural resources currently have relatively few approaches that can search specific terms at different temporal spans and semantic relationships between terms that may enable significant events and event patterns to be identified. In addition to the visual and quantitative analyses provided within AIDA that facilitate event identification based on when they are reported, our approach provides statistical text output that can be exported to analytical tools (e.g., R) for further analysis. Because of these capabilities, our presented approach, therefore, is unique in its applicability to monitoring land use events at different regional scales.

Our intention is to apply the techniques presented here to settings that are not as highly monitored as states in the Midwest. We intend to enable and apply web content that can be continuously observed and searched using RSS feeds or other electronic formats using the same approach presented in the case study. Such collected data can be observations made by local populations and analyzed using the techniques presented in order to better understand land use change issues that are affected by a variety of social and ecological factors. This enables the analysis to look at either singular events or long-term event patterns as described by individuals. The application of AIDA for real time analysis of social data and other observations would be particularly useful for settings where long-term analysis is necessary but the collection of data over long periods is difficult. In addition, both structured and unstructured formats can be used, enabling GIS and network analysis to be coupled. For specific land use topics, terms relevant in observations should be added to publically accessible databases; this enables future searches to have richer and more comprehensive sets of terms. With regard to improving our search techniques, we do envision applying more sophisticated artificial intelligence and learning techniques applied by other search software in future versions of AIDA [37,38]. Regardless of the analysis technique, because our approach allows term searches to be conducted at different time scales along with the identification of semantic relationships, we believe these general principles presented here can have significant benefit for regions that lack land use monitoring, particularly observations concerning relevant social processes that are difficult to physically measure.

One area that our approach can add significant benefit is in social-ecological modeling, including the use of agent-based [39] modeling and other land use modeling techniques. Currently, many models that integrate empirical social-ecological data suffer from a lack of relevant information at various time and spatial scales [40]. This lack of data for models, particularly those that incorporate agent-based techniques, has often led to models that cannot be easily validated or mapped to the real world [41], limiting such models for forecasting purposes that attempt to understand how land use change may

affect specific settings. The integration of web-based data collection and data mining tools can facilitate more rapid and cost-effective techniques in accumulating information. Our intent is that the data collection methods and analysis presented here can be directly streamed into existing or new models so that these models are parameterized and calibrated based on continuous data over time. For example, specific term patterns and relationships associated with land use change can be used to inform models on the likelihood of critical decisions being made. These trends and behaviors can be translated into probabilities or parameters for specific decisions (e.g., policy concerning grassland conservation) based on pertinent data captured (e.g., legislative details, budgetary considerations, *etc*.). In essence, such an approach may one day function similarly to environmental modeling (e.g., hydrologic or weather modeling) in that data streams can be used to continuously calibrate and adjust models [42] as needed in order to enable models to better forecast trends and explain social-ecological dynamics affecting land use. Therefore, in addition to applying AIDA and such tools to monitoring, we believe the methods and tools presented here should be integrated into existing modeling methods and tools. We plan to focus on this endeavor in future efforts.

Acknowledgements

We are grateful to the National Science Foundation (OPP Arctic System Science #0531148 and #0755966, and Experimental Program to Stimulate Competitive Research #0701898 and #0919608) for funding this research. The views expressed here do not necessarily reflect those of the National Science Foundation.

References and Notes

1. Pielke, R.A. Land use and climate change. *Science* **2005**, *310*, 1625-1626.
2. Pirot, J.Y.; Meynell, P.; Elder, D. *Ecosystem Management: Lessons from Around the World*, 1st ed.; IUCN: Gland, Switzerland, 2000.
3. Folke, C.; Hahn, T.; Olsson, P.; Norberg, J. Adaptive governance of social-ecological systems. *Annu. Rev. Environ. Resour.* **2005**, *30*, 441-473.
4. Guralnick, R.; Hill, A. Biodiversity informatics: Automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics* **2009**, *25*, 421-428.
5. Altaweel, M.; Alessa, L.; Kliskey, A. Visualizing situational data: Applying information fusion for detecting social-ecological events. *Soc. Sci. Comp. Rev.* **2010**, doi:10.1177/0894439309360837.
6. Zhang, J.; Gruenwal, L.; Gertz, M. VDM-RS: A visual data mining system for exploring and classifying remotely sensed images. *Comput. Geosci.* **2009**, *35*, 1827-1836.
7. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*, 2nd ed.; Sage: Thousand Oaks, CA, USA, 2004.
8. Landmann, J.; Zuell, C. Identifying events using computer-assisted text analysis. *Soc. Sci. Comp. Rev.* **2008**, *26*, 483-497.
9. Dumouchel, W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous report system. *Am. Stat.* **1999**, *53*, 177-190.

10. Hand, D.J. Pattern detection and discovery. In *Pattern Detection and Discovery*, 1st ed.; Lecture Notes in Artificial Intelligence; Hand, D., Adams, N., Bolton, R., Eds.; Springer-Verlag: New York, NY, USA, 2002; Volume 2447, pp. 1-12.
11. Pharo, N.; Järvelin, K. The SST method: A tool for analysing Web information search processes. *Inform. Process. Manag.* **2004**, *40*, 633-654.
12. Rushing, J.; Ramachandran, R.; Nair, U.; Graves, S.; Welch, R.; Hong, L. ADaM: A data mining toolkit for scientists and engineers. *Comput. Geosci.* **2005**, *31*, 607-618.
13. Wu, S.Y.; Chen, Y.L. Mining nonambiguous temporal patterns for interval-based events. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 742-758.
14. Smith, G. *Illinois Wind Turbines: Florida Company Wants to Build 133 in Dekalb County*, *18 in Lee County*; Chicago Tribune: Chicago, IL, USA, 3 April 2009; Available online: http://archives.chicagotribune.com/2009/apr/03/local/chi-wind-farm-debate-03-apr03 (accessed on 26 December 2009).
15. Walker, B.; Holling, C.S.; Carpenter, S.R.; Kinzig, A. Resilience, adaptability, and transformability in social-ecological systems. *Ecol. Soc.* **2004**, *9*, 5; Available online: http://www.ecologyandsociety.org/vol9/iss2/art5/ (accessed on 26 December 2009).
16. Arens, Y.; Knoblock, C.A.; Shen, W.M. Query reformulation for dynamic information integration. *J. Intell. Inf. Syst.* **1996**, *6*, 99-130.
17. Torra, V. *Information Fusion in Data Mining: Studies in Fuzziness and Soft Computing*, 1st ed.; Springer-Verlag: Berlin, Germany, 2003.
18. Mitra, P.; Wiederhold, G.; Kersten, M. A graph-oriented model for articulation of ontology interdependencies. In *Proceedings of the 7th International Conference on Extending Database Technology: Advances in Database Technology (EDBT 2000)*, Konstanz, Germany, March 2000; pp. 86-100.
19. Zhai, Y.; Shah, M. Tracking news stories across different sources. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, Singapore, November 2005; pp. 2-10.
20. *Google Trends* Homepage. http://www.google.com/trends (accessed on 27 December 2009).
21. Sallach, D.L.; Jozik, J. Data theory, discourse mining and thresholds. In *Proceedings of the Complex Adaptive Systems and the Threshold Effect: Views from the Natural and Social Sciences: Papers from the AAAI Fall Symposium*, Arlington, VA, USA, November 2009; pp. 110-116.
22. *AIDA*. Downloadable code and example can be obtained in the repository of *Sustainability* at: http://www.mdpi.com/2071-1050/2/5/1182/s1.
23. *UIMA* Homepage. http://incubator.apache.org/uima/ (accessed on 28 December 2009).
24. *GATE* Homepage. http://gate.ac.uk (accessed on 28 December 2009).
25. *WordNet Search* Homepage. http://wordnetweb.princeton.edu/perl/webwn (accessed on 31 January 2010).
26. *TreeTagger* Homepage. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ (accessed on 29 December 2009).
27. Gospondnetic, O.; McCandless, M. *Lucene in Action*, 2nd ed.; Manning Publications: Greenwhich, CT, USA, 2005.
28. *Lucene* Homepage. http://lucene.apache.org/java/docs/ (accessed on 29 December 2009).

29. Salton, G.; Buckley, C. Weighting approaches in automatic text retrieval. *Inform. Process. Manag.* **1988**, *24*, 513-523.
30. *GeoTools* Homepage. http://www.geotools.org/ (accessed on 21 February 2010).
31. Alessa, L.N.; Kliskey, A.D.; Williams, P.; Barton, M. Perceptions of change in freshwater remote resource-dependent Arctic communities. *Global Environ. Chang.* **2008**, *18*, 153-164.
32. Speaker to address the "greening" of Lincoln. *Lincoln Journal Star*, 3 September 2009; Available online: http://journalstar.com/news/local/article_da526d18-980c-11de-b2d9-001cc4c03286.html (accessed on 23 February 2010).
33. Heitz, D. Senior housing project in Bettendorf moves forward. *Quad-City Times*, 2 June 2009; Available online: http://www.qctimes.com/news/local/article_46e08238-4fda-11de-960d-001cc4c002e0.html (accessed on 23 February 2010).
34. Grassland bird study could help population. *Journal Star*, 15 June 2009; Available online: http://journalstar.com/news/local/article_6c73ef9a-db6e-523f-8deb-52ecbc03cfaf.html (accessed on 23 February 2010).
35. Federal money available to preserve grasslands. *Journal Star*, 14 June 2009; Available online: http://journalstar.com/news/state-and-regional/govt-and-politics/article_b10968b8-bbec-5b18-b542-ed40ec2ef973.html (accessed on 23 February 2010).
36. U.S. Government Printing Office. *Public Law 111-5-American Recovery and Reinvestment Act of 2009*; Available online: http://www.gpo.gov/fdsys/pkg/PLAW-111publ5/content-detail.html (accessed on 19 April 2010).
37. Bunescu, R.C.; Mooney, R.J. Extracting relations from text: From word sequences to dependency paths. In *Natural Language Processing and Text Mining*, 1st ed.; Kao, A., Poteet, S.R., Eds.; Springer-Verlage: London, UK, 2006; pp. 29-44.
38. Chambers, N.; Jurafsky, D. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Singapore, August 2009; Volume 2, pp. 602-610.
39. Bonabeau, E. Agent-based modeling: Methods and techniques for simulating human systems. *Proc. Nat. Acad. Sci. USA* **2002**, *99*, 7280-7287.
40. Altaweel, M.; Alessa, L.; Kliskey, A.; Bone, C. A framework to structure agent-based modeling data for social-ecological systems. *Struct. Dynam. eJ. Anthro. Rel. Sci.* 2010, (in press).
41. North, M.J.; Macal, C. *Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation*, 1st ed.; Oxford University Press: New York, NY, USA, 2007.
42. Gan, T.Y.; Dlamini, E.M.; Biftu, G.F. Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling. *J. Hydrol.* **1997**, *192*, 81-103.