Ecosystems Monitoring: An Information Extraction and Event Processing Scientific Workflow

Ahmed Abdeen Hamed and Byung Suk Lee
Department of Computer Science
University of Vermont, UVM
Burlington, Vermont 05405
Email: {ahamed, bslee}@cems.uvm.edu

Anne E. Thessen
Encyclopedia of Life
Marine Biological Laboratory, MBL
Woods Hole, Massachusetts 02543
Email: athessen@mbl.edu

***Abstract*—This paper presents a novel architecture that brings together Information Extraction (IE) with Event Processing (EP) research areas to globally monitor human activities and biodiversity dynamics and measure their impact on ecosystems. The two areas (IE and EP) are rich on their own and we believe their integration will achieve a much more comprehensive solution to ecosystems monitoring. The integration is based on a closed-loop mechanism that guarantees the communication and the evolution of the overall architecture. While we use Microblogging communities (e.g., Twitter) as a news producing tool, we keep track of the vulnerable ecosystems using a GIS tracking database. We also make use of Google Map/Earth API capabilities to dynamically update the GIS database. After a complete cycle, the architecture produces a list of vulnerable ecosystems. This architecture leverages the rich research in Scientific Workflows to achieve the integration and communication of the various components. We are in the process of developing a system that can be used by conservationists and decision makers to efficiently allocate their time and limited resources in response to ecosystems perturbation.**

I. Introduction

The effects of human activities on ecosystems and biodiversity have increased so much that the rate of species extinction is rising hundreds or thousands of times [2]. Biodiversity is important not only because of its intrinsic value but also because of the basic services it provides, without which humans could not survive [26]. With this in mind, this paper presents a novel monitoring approach that focuses on ecosystems that suffer destructive activities, by humans or the ecosystem's living species.

The approach has the following features: (i) integration of Event Processing and Information Extraction research areas in a closed-loop, (ii) dynamic monitoring of the biodiversity ecosystems and related human activities, and (iii) the use of Twitter as a news broadcasting tool and the main source of events that impacts ecosystems. The paper formalizes this approach using Scientific Workflows to describe the scientific aspects, integrate the various components, and capture and analyze data in each stage [13]. This formalization results in a complex architecture that captures destructive activities, quantifies their impact on an ecosystem, and based on this expresses vulnerability scores for that ecosystem. Each vulnerability score associated with an ecosystem corresponds to a classification category. We envision this architecture as a flexible, experimental platform where various combinations of data sets and algorithms are plugged-in until the best results are achieved. We are currently using KeplerWeka to prototype each component of the workflow before assembling it. KeplerWeka, as the name suggests, is a hybrid workflow modeling environment that integrates Kepler [31] with Weka [32]. While Kepler is specialized in pure workflow modeling, Weka is specialized in data mining and machine learning algorithms which will benefit our project a great deal. Weka comes with a wide varity of built-in algorithms and can be plugged with different parameters to come up with the best results. [30]

This paper is structured as follows: Section II provides the necessary background on Information Extraction, Event Processing, Microblogging communities and Ecosystems Dynamics. Section III provides the overall architecture, components and data flow. In section IV the paper discusses the immediate future directions and conclusions.

II. Background

A. Information Extraction Background

Information Extraction is a technique used to detect relevant information in a large set of unstructured documents and present them in a machine readable format. This technique is used to analyze the text to locate certain pieces of information in the text [12].

There are two main approaches to designing an information extraction system: *knowledge engineering* and *automatic training* [5]. The knowledge engineering approach uses handcrafted grammars expressing the rules of the application domain knowledge, and the automatic training approach uses machine learning techniques on a training set developed by a domain expert.

The dramatic growth in the number of textual resources available on the web has lead to increased interest in information extraction [5]. The accomplishments made include online databases like PubMed which offers access to the ever-increasing MEDLINE [34] dataset. In addition to online databases, Really Simple Syndication (RSS) [33] enables users to subscribe to their choice of content sources across the web. Aggregation tools (e.g., RSS readers) display summaries of these subscriptions, which update automatically when new information is available [7]. Moreover, online resources have

become an attractive source of information due to the massive content their users publish everyday.

In general, Information Extraction plays a very important role in processing the contents of various online resources and enables the science behind all those available resources. In particular, Information Extraction in our project is concerned with identifying the twitter's ongoing theme(e.g., oil spill, deforestation, dam development). It also goes deeper to the article cited by the tweet or the RSS feed and explores the activity aspects. For each activity, basic knowledge is extracted (i.e., geolocation of the activity, species affected, area of the impact).

B. Event Processing Background

An event refers to any occurrence of a phenomenon, data acquisition, a notification, etc., which triggers a reaction. Triggering a reaction may require one or more events, and the outcome of a reaction can be to create one or more other events, each of which may in turn triggers another reaction. A reaction can involve any number of operations, such as reading the input that triggered the reaction, transforming the input to something else, creating a new event as an output, and deleting (or ignoring) irrelevant events.

The objective of event processing varies significantly depending on the application domain. (e.g., to change the behavior of the system dynamically in order to react to incoming events, to look for exceptional behavior and generate alerts when such behavior occurs, to deliver the right information to the right consumer in the right granularity at the right time, to diagnose a problem, based on observed symptoms to identify events before they happen, so that they can be eliminated or at least have their effects mitigated)[6].

Event processing in our project can be viewed as the problem of updating the vulnerability score of each ecosystem monitored in an ecological network. The vulnerability scores of all ecosystems collectively define the state of the entire network, and the number of possible states of the network can be very large (e.g., 5^{100000} with 5 vulnerability scores and 100000 ecosystems). However, the actual updates are typically limited to a portion of the ecosystems in the network. Evidently, effective interactions between IE and EP are important to identify the portion correctly in the dynamic ecosystem.

C. Microblogging Communities

Microblogging is a form of communication that takes place on an online social network by whereby users broadcast brief text updates, also known as tweets, to the public [21]. A recent analysis of the Twitter network revealed uses of microblogging for news reporting, (e.g., commentary on news and current affairs [20] [21]). Twitter has been used for various applications: for example (i) to measure the happiness level in written expressions from a large collection of twitters [22] or (ii) to track the stock market where a specialized list of twitters broadcast the ticker prices continuously as they fluctuate [29].

Similarly, this workflow uses Twitter [28] as a monitoring tool that keeps track of any perturbation activities (e.g., land transformation due to human activities) that can possibly degrade, transform or destroy an ecosystem. The workflow will monitor biodiversity news broadcasters that have a high level of credibility(e.g., Eco Conservation, IUCN, Nature, CNN, WWF, Ocean Defense, PBS Nature, Wild Life, Discovery, Science Daily, BBC News). Those are the kinds of microbloggings that we start with. However, personal tweets are also considered when an expression of opinion is broadcasted such as expressed disapproval of a highway built on the expense of destroying a large area of land (e.g., the ever debated Indiana I-69 Interstate).

D. Ecosystems Dynamics

An ecosystem can be described as a collection of species, their shared habitat and their interactions. Part of what makes ecosystem analysis so complex is the broad variation in the types and strength of these interactions. For example, organisms can have predator-prey relationships, parasite-host relationships, competitive relationships or commensal relationships. Some relationships can change over time, while others are static. Any successful attempt to model an ecosystem must be able to capture the dynamic nature of these relationships as a network analysis problem. The concept of using network analysis in ecology is not new [19]. What is new is the development of a computational methodology specifically designed to analyze and predict complex networks. Ecological networks can be applied to answer questions about conservation or restoration ecology and manipulate ecosystems and risk assessment [11]. We use a GIS database to keep track of the ecological networks as discussed in III-C1.

III. Overall Architecture

The scientific workflow components and interactions among the components are demonstrated in Figure 1

A. A Monitoring Subworkflow

This subworkflow is for monitoring conservation and biodiversity news broadcasted via preselected reliable Twitter lists (e.g., Encyclopedia of Life and Biological Science). The monitor's role, besides capturing the tweets contents, is to analyze the content and classify them as relevant or irrelevant to a particular ecosystem. The monitor employs a feature selection algorithm (e.g., genetic algorithm) [18] to filter putative features before using them in a binary classification algorithm (e.g, immune or Bayesian classifier) [8]. If a tweet is classified to be relevant, the monitor sends the tweet to the Information Extraction subworkflow. Otherwise, no action is taken and the tweet is ignored.

B. An Information Extraction Subworkflow

When a tweet is identified as relevant it is delivered to the Information Extraction subworkflow for further analysis to capture the destructive events and their impact on the ecosystem. Most tweets have embedded links (i.e., URL) in their bodies that refer to the source of news broadcasted. This URL must be visited and its content must also be

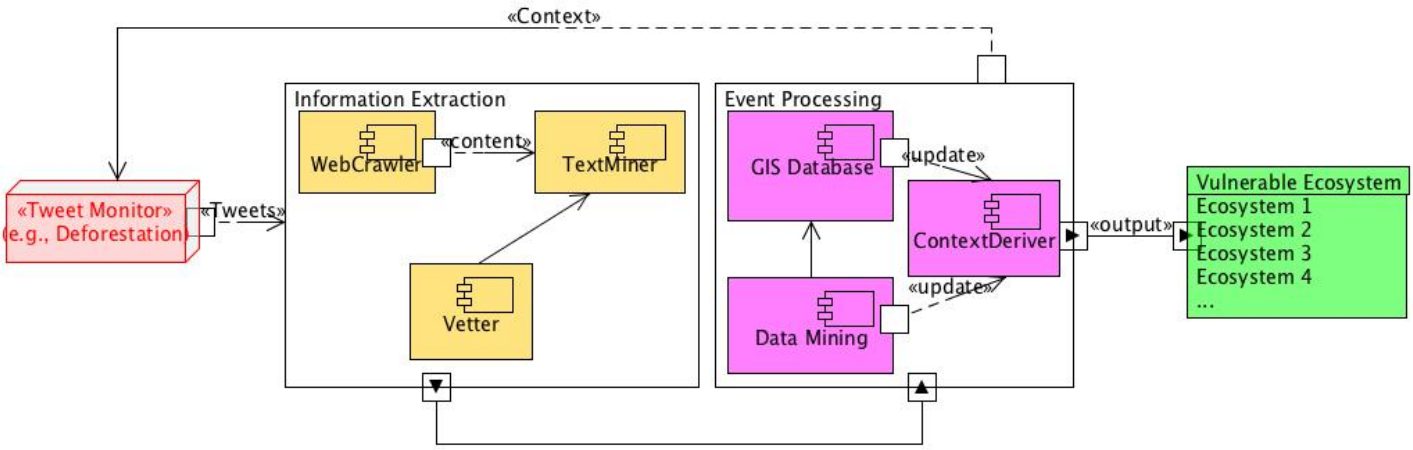

Fig. 1. Ecosystems Monitoring Scientific Workflow

analyzed. Therefore, the Information Extraction subworkflow is comprised of the following subcomponents:

1) WebCraweler: This subcomponent's role is to visit the URLs embedded in the body of the tweet. A web crawler may also crawl to a deeper level into other URLs referred to by the article. When the source article is visited its contents are delivered to the TextMiner subcomponent to extract all related data items. We use Apache Lucene for the purposes of keyword indexing and constructing a hyperlink graph [25].

2) TextMiner: When the article contents are delivered by the WebCrawler, the TextMiner performs the various tasks in a pipeline fashion. First, the article contents must be stemmed to eliminate stopwords and any noisy words. Second, a sentence detector must be applied to identify relevant pieces of information. Third, a sentence tokenzier must be applied on the tokens level to identify what entity each token represents (e.g, geolocation, organism, environment, organization, etc). Fourth, we apply a part-of-speech detector to identify nouns, proper nouns and verbs so we can identify relationships between entities (e.g., the impact of an oil spill on the Gulf of Maine). There is a good variety of text mining tools that provide a standalone analysis engine for each component of the pipeline. We are utilizing open-sources frameworks that are widely accepted by the text mining community (e.g., UIMA and GATE) [24] [23].

3) DataVetter: When the pieces of data items are extracted they are vetted and each item must be verified before we measure the impact of destructive events. This step can be either done manually by human experts where the items are queued for final approval or can also be automated and done via a verification software subcomponent. For the system that we are currently developing this step is automated using a supervised machine learning approach. However, the training set will be identified manually be domain experts. When data items are verified, they become ready for processing by the Event Processing subworkflow and hence pushed forward.

C. An Event Processing Subworkflow

When all the pieces of information about a certain event (e.g., land transformation, degradation or deforestation activities) are gathered they are used to estimate the impact on ecosystems. This subworkflow is comprised of the following subcomponents:

1) GIS Tracking Database: The GIS database keeps track of the vulnerable ecosystems geographically and serves as a mechanism that delivers the impact to other ecosystem. Ecosystems can be naturally envisioned as a complex network where the nodes represent the individual ecosystems and the edges represent the relationships between one ecosystem and another (e.g., neighbor_of, competes_with, adjacent_to). Thus, impact may propagate from one ecosystem to another connected ecosystem via a relationship. As a result, the neighbors of an ecosystem (e.g., other adjacent ecosystems) may be affected and become vulnerable as well. The network of ecosystems will be constantly analyzed to derive a new context. We are utilizing the World Wild Foundation (WWF) Terrestrial Ecoregions GIS Database [27] for this particular task. We are also leveraging the open-source GeoTools API to query and browse the WWF database and we use Google Maps and Google Earth API to capture the geographical features of an ecoregion.

2) DataMiner: Data items that are extracted by the Information Extraction subworkflow are used in computing the impact of a particular event (e.g., over fishing, habitat transformation, pollution activities). This subcomponent is designed to analyze the data items and identify which ecosystems are affected and quantify the impact. Data mining and machine learning techniques will be used to measure the impact. Manual microblogs (i.e., tweets) are currently selected in the process of building a training set. The DataMiner

subcomponent may also utilize data produced by an ecosystem simulation environment to reflect the dynamic changes on the ecosystem. This results in quantifying the impact of the activity. In the process of quantification, the DataMiner may need to request further information about the ecosystem under the impact which could lead to getting it back from GIS database via the Google Earth APIs.

3) ContextDeriver: Due to the score changes of impacted ecosystems, other GIS nodes (ecosystems) may change as well. Hence, the scores of the some of its neighbors may change and must be recalculated to reflect the new dynamics. This leads to the reclassification of the ecosystem's vulnerability level and the emergence of new vulnerable ecosystems which must be monitored. This subcomponent must traverse the GIS database after each update operation to identify the new vulnerable ecosystems. Various graph search algorithms (e.g., Depth-First Search) can be used to identify those ecosystems. When the ContextDeriver identifies the target ecosystems, it notifies the monitor and hence the monitor "tunes" itself to listen to activity that pertains to the newly identified ecosystems. With this step accomplished, this completes the proposed scientific workflow for ecosystem vulnerability assessment and the system produces an ordered list of all vulnerable ecosystems.

IV. Future Directions and Conclusion

The paper describes the ongoing development of our system. We have successfully modeled the Monitoring Subworkflow. We also used J48 for a multi-class classification algorithm as a starting point. In the future we will use other classification algorithms (e.g., adaptive immune classifier or Bayesian classifier). We have also implemented the GIS database and extracted the graph of Ecoregions for analysis. The database will be extended to a Spatio-temporal Semantic Database to keep track of the destructive event and their various aspects. This will enable time-oriented queries to be made against the database (e.g., queries for a certain event type occurring in a given duration of time, queries for various event types in a given duration of time taken place in a particular ecoregion). We anticipate many computational problems to emerge (e.g. workflow adaptability, components mapping, and provenenace), and they will be tackled accordingly.

Acknowledgment

We would like to thank David Patterson and Peter Mangiafico of the EOL, Holly Miller of the MBL, Ferdinando Villa of the Gund Institute and Gary Johnson of UVM for their valuable discussions.

References

[1] R. Bendall. Biodiversity: the follow up to rio. The Globe, April 1996.
[2] P. Berry. Global biodiversity assessment, unep : edited by vh heywood cambridge university press, cambridge, 1996. *Global Environmental Change*, 7(2):191 – 192, 1997.
[3] G. E. Change. Global environmental change: Human and policy implications. *Population, Land Management and Environmental Change*, September 1995.
[4] Delaware, Maryland, N. Jersey, Pennsylvania, W. Virginia, and USDA. Pollination. *MAARC Publication*, 5.2, 2000.
[5] L. Eikvil. Information extraction from world wide web - a survey. Technical Report 945, Norweigan Computing Center, 1999.
[6] O. Etzion and P. Niblett. *Event Processing in Action*. Manning Publications Co., 2009.
[7] J. Grossnickle. Rss - crossing into the mainstream. White paper, Yahoo, October 2005. Available online (12 pages).
[8] D. Hernández-Lobato and J. M. Hernández-Lobato. Bayes machines for binary classification. *Pattern Recogn. Lett.*, 29(10):1466–1473, 2008.
[9] V. H. Heywood. *Global biodiversity assessment*. Cambridge University Press, October 1995.
[10] V. H. Heywood. The global biodiversity assessment. The Globe, April 1996.
[11] J. Memmott The restoration of ecological interactions: plantpollinator networks on ancient and restored heathlands Journal of Applied Ecology, 2009.
[12] K. Kaiser and S. Miksch. Information extraction. a survey. Technical Report Asgaard-TR-2005-6, Vienna University of Technology, Institute of Software Technology and Interactive Systems, 2005.
[13] Z. Lacroix, C. Legendre, and S. Tuzmen. Reasoning on scientific workflows. In *SERVICES I*, pages 306–313, 2009.
[14] I. Muslea. Extraction patterns for information extraction tasks: A survey. In *In AAAI-99 Workshop on Machine Learning for Information Extraction*, pages 1–6, 1999.
[15] Standards and Petitions Working Group of the IUCN SSC Biodiversity Assessments Sub-Committee. Guidelines for using the IUCN red list categories and criteria. http://intranet.iucn.org/webfiles/doc/SSC/RedList/RedListGuidelines.pdf, August 2008.
[16] J. I. Uitto and A. Ono. Population, land management and environmental change. *The United Nations University*, 1996.
[17] USFWS. U.s. fish and wildlife service report to congress. News release, July 1994.
[18] H. Vafaie and K. D. Jong. Genetic algorithms as a tool for feature selection in machine learning. In *in Machine Learning. In Proceedings of the 1992 IEEE Int. Conf. on Tools with AI*, pages 200–204. Society Press, 1992.
[19] B. M. Hannon. Structure of ecosystems. *Journal of Theoretical Biology*, 41:697– 706, 1973.
[20] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.
[21] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*, abs/0911.1583, 2009.
[22] P. Dodds and C. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 2009.
[23] K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham. Evolving gate to meet new challenges in language engineering. *Natural Language Engineering*, 10(3/4):349—373, 2004.
[24] Apache Software Foundation. Apache UIMA.
[25] Apache Software Foundation. Apache Lucene.
[26] R. Costanza, d'Arge Ralph, R. de Groot, S. Farber, M. Grasso, B. Hannon, K. Limburg, S. Naeem, R. V. O'Neilland, and J. Paruelo. The value of the world's ecosystem services and natural capital. *Ecological Economics*, 25(1):3–15, 1998.
[27] David M. Olson, Eric Dinerstein, Eric D. Wikramanayake, Neil D. Burgess, George V. N. Powell, Emma C. Underwood, Jennifer A. Doamico, Illanga Itoua, Holly E. Strand, John C. Morrison, Colby J. Loucks, Thomas F. Allnutt, Taylor H. Ricketts, Yumiko Kura, John F. Lamoreux, Wesley W. Wettengel, Prashant Hedao, and Kenneth R. Kassem. Terrestrial ecoregions of the world: A new map of life on earth. *Bioscience*, 51(11):933–938, 2001.
[28] http://twitter.com
[29] http://twitter.com/Stockonews
[30] http://sourceforge.net/projects/keplerweka/
[31] https://kepler-project.org/
[32] http://www.cs.waikato.ac.nz/ml/weka/
[33] http://web.resource.org/rss/1.0/
[34] http://www.ncbi.nlm.nih.gov/pubmed/