To appear in *Proceedings of The 17th International Conference on Data Engineering,* April 2-6, 2001, IEEE.

GeoNODE: An End-to-End System from Research Components

Chris Clifton
clifton@mitre.org

John Griffith
jgriffith@mitre.org

Rod Holland
holland@mitre.org

The MITRE Corporation
202 Burlington Road
Bedford, MA 01730-1420 USA

Abstract

MITRE's GeoNODE (Geographic News On DEmand) project has used the results of several research efforts to build an end-to-end "live" prototype: a system that addresses all the issues necessary to tie research ideas into a real infrastructure. GeoNODE incorporates concepts, products, and research prototypes from the database, data mining, semistructured data, multimedia, natural language, visualization, and geographic information system communities to build an end-to-end analysis tool for collections of news. We will demonstrate the GeoNODE system, and give an overview of the research components used to create it. We will discuss the lessons learned, particularly issues in infrastructure "awareness" that affect future research projects, and benefits to our research from participating in this project.

Putting research into practice can be difficult. Occasionally research results translate directly into a stand-alone product, but more often the research developments are only useful in the context of a larger system. The research literature is full of brilliant ideas that have never been used: The cost of integrating those ideas into an end-to-end system was too high. The result is that research is left by the wayside: Potential solutions are ignored in favor of technically inferior, but more easily integrated, alternatives. A grand vision isn't enough (even if the hard research issues are solved): End users are more impressed with a demonstrable system that solves their needs.

MITRE has conducted considerable research into technologies for text and multimedia analysis. These projects were in danger of falling into the "wonderful vision" category: Although implemented and demonstrable by themselves, no single project satisfied and end-user need.

The GeoNODE project was started to address this: The goal was to combine these technologies into a complete, usable prototype. GeoNODE incorporates elements of multimedia processing, natural language processing (NLP), semistructured data, database, data mining, information retrieval, geographic information systems (GIS), and visualization; under a specific application profile: that of a tool supporting mixed-initiative intelligence analysis. A bumper-sticker summary of this capability is "news on a map". (MITRE's charter doesn't include producing *products* – GeoNODE is a prototype that proves 1) a product could be built, and 2) such a product would fill a real need.) GeoNODE has succeeded: We have users who are convinced it represents a valuable system, and want to test the prototype on their own problems.

Whether such an effort constitutes research in itself, or advanced development, is a matter of semantics. Our motives for building such a system were several:

- To concretely demonstrate the promise implicit in a number of new technologies by producing a capability that did not hitherto exist;
- To demonstrate synergies among previously unrelated information technologies;
- To inform multiple interested communities that such systems can now be built, thereby stimulating the transfer of research technologies to real-world users;
- To explore the capabilities, properties, and problems of systems of this class;
- To provide researchers with an integrated framework in which to exercise their components, perform experiments, and rapidly improve.

We have learned many lessons along the way. Making these components work together – turning individual research prototypes into a cooperative system – has not been easy. Sometimes the difficulties have been just engineering

"glue", but we have also identified harder challenges that have turned into research problems in their own right.

This demonstration will show what GeoNODE does and how it works. We will describe the individual components and how they play together. We will discuss the lessons learned, particularly issues in infrastructure "awareness" that affect future research projects. We will conclude with suggestions on how other research projects can benefit from building end-to-end prototypes, and the costs of doing so.

1. What GeoNODE Does

GeoNODE allows a human analyst to perform research and analysis against multiple real-world information sources – e.g., broadcast news and web-hosted newspapers – to perform what we term "mixed initiative intelligence analysis". As in other mixed initiative systems, some actions take place at the initiative of the human user (e.g., submission of a query), some actions take place at the initiative of the system (e.g., automatic nomination of a topic cluster). For example, the user may formulate a query meant to find stories dealing with the proliferation of weapons of mass destruction; the system may independently nominate a topic cluster consisting of stories having to do with a round of Indo-Pakistani nuclear weapons tests. In general, user initiative affordances allow expression of the user's will; system initiative capabilities inform or alert the user to phenomena in the information stream which may require further attention.

GeoNODE combines NLP and data mining processing of information sources that are essentially linguistic (the news) with GIS-based information fusion techniques which had previously been applied only to signal processing-based sources. This leads to a capability to automatically trace the evolution of the news in space and time: news on a map. What follows is a brief summary of processing in the GeoNODE system; aspects of this processing are discussed in detail later.

Real-world information sources (broadcast news[4] and web-hosted newspapers) are collected daily, for research purposes. For each source, the raw data is segmented (or "zoned") into their constituent stories[1]. On any given day, from 8-15 sources are being collected and segmented.

Each story is processed by MITRE's Alembic system [3] to extract named entities as metadata. These are the names of persons, places and organizations – identified as such – found in the story. For each story, place names are assigned geospatial coordinates (latitude and longitude) by reference to a gazetteer; these geo-referenced place names further enrich the metadata. Story datelines are used as a source of temporal metadata.

The stories plus the extracted metadata comprise a database for further processing. The named entities serve as a term space for subsequent correlation and cluster steps (the TopCat[2] pipeline), yielding a collection of automatically nominated (system-initiative) topic clusters. A geographic information system (ESRI's ArcView™) supports queries against the spatiotemporal metadata. Traditional SQL queries are also supported.

1.1 A GeoNODE Example

An example of a GeoNODE session is shown in Figure 1. The source is CNN; two related topic clusters have been nominated which describe the bombing of the US embassies in Kenya and Tanzania, and the subsequent US cruise missile strikes on Sudan and Afghanistan. Topic clusters are expressed as lists of named entities. The cluster describing the embassy bombings is designated Afghanistan, Kenya, Nairobi, Tanzania; the cluster describing the US counterstrike is Afghanistan, Khartoum, Ladin, Sudan.

These two topics (among many) are highlighted in the Topic listbox in a portion of the GeoNODE interface (shown here in the lower right hand corner of the figure, partially obscuring the map display) which controls views of the CNN source. The time interval for the view of these topics has been constrained to the period between May 2, 1998 and September 30, 1998. A histogram view of the selected topics, plotting number of stories on topic against time, is shown in the lower left-hand corner of the figure.

In the map display in the upper half of the figure, a World View window shows a global summary of all place names mentioned in the selected topics. A colored circle is shown for each named location; the diameter of a circle is an indicator of frequency of mention. The color corresponds to the source of the story.

The left hand map display presents a Zoom View of a region of interest (East Africa, Europe, and Central Asia). This display again shows the place names mentioned in stories on the selected topics. The demonstration will show other features, such as an animation of locations mentioned in topics over time. The animation may be run from beginning to end, or subject to direct temporal manipulation by the user. This view of the selected topics shows the evolution of events, from the initial bombing and the world reaction, to the US cruise missile strikes which followed (and the world reaction to them), and residual consequences (to the extent to which they were reported by CNN).

Acknowledgements

We would like to thank the following people whose work and contributions have been necessary to the GeoNODE project: Tom Bartee, Stanley Boykin, Thad Cooper, Shereif El-Sheikh, Gerry Fitzgerald, Joe Francoeur, John Gibson, Steve Hansen, Lynette Hirschman, Rob Hyland, Steve Ja-

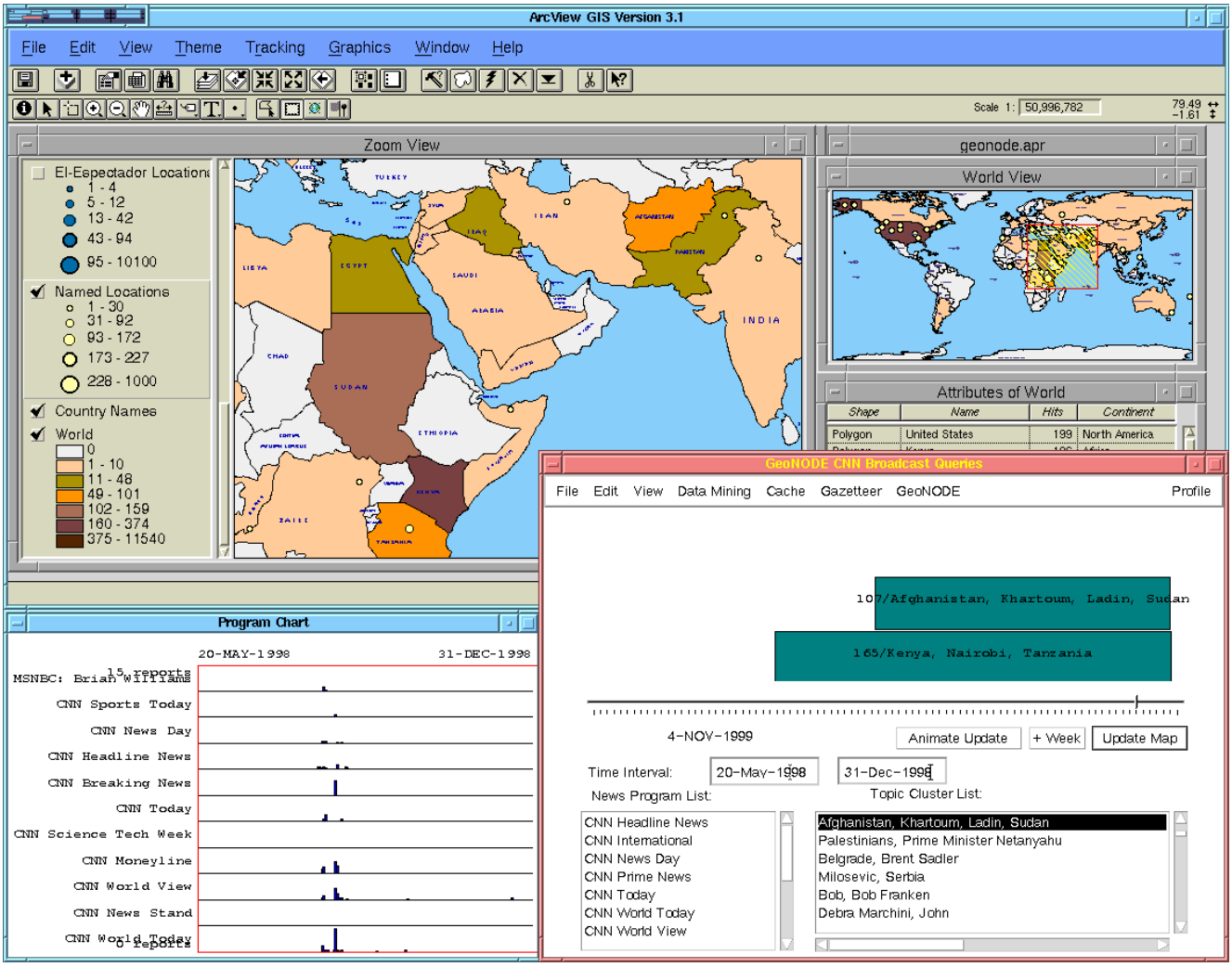

Figure 1. GeoNODE screen shot.

niak, Mark Maybury, Andy Merlino, Carsten Oertel, Marc Reichman, Marc Richards, Paul Silvey and Shane Steward.

References

[1] S. Boykin and A. Merlino. Machine learning of event segmentation for news on demand. *Commun. ACM*, 43(2):35–41, Feb. 2000.

[2] C. Clifton and R. Cooley. Topcat: Data mining for topic identification in a text corpus. In *3rd European Conference on Principles and Practice of Knowledge Discovery in Databases*, Prague, Czech Republic, Sept. 15–18 1999.

[3] D. Day, J. Aberdeen, L. Hirschman, R. Kozierok, P. Robinson, and M. Vilain. Mixed initiative development of language processing systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D.C., Mar. 1997.

[4] M. Maybury, A. Merlino, and D. Morey. Broadcast news navigation using story segments. In *ACM International Multimedia Conference*, Seattle, WA, Nov. 1997.