Automated Knowledge Annotation for Dynamic Collaborative Environments

Andrew J. Cowell, Michelle L. Gregory, Eric J. Marshall, Liam McGrath
Pacific Northwest National Laboratory
PO Box 999, Richland, WA, 99352, USA
{andrew | michelle | eric.marshall | liam.mcgrath}@pnl.gov

Abstract

This poster describes the Knowledge Encapsulation Framework (KEF), a suite of tools to enable automated knowledge annotation for modeling and simulation projects. This framework can be used to capture evidence (e.g., facts extracted from journal articles and government reports), discover new evidence (from similar peer-reviewed material as well as social media), enable discussions surrounding domain-specific topics and provide automatically generated semantic annotations for improved corpus investigation.

Introduction

Researchers across all domains in academia, industry and government have the onerous task of keeping up with literature in their fields of study. The use of the Internet has made long distance collaborations possible and thus has increased productivity of researchers in general. In addition, the Internet makes it easier for academic journals, conferences, workshops, and individual researchers to put their content in front of a larger audience. It has also made it easier than ever to perform searches and find relevant information.

However, the use of the Internet as a research tool has its limitations due to the quantities of data available and often questionable quality (not to mention the multitude of file formats and standards). Researchers are finding it more difficult to identify relevance and significance of individual articles in the mass of similarly titled material. Once material is found, the benefits of electronic media end there: researchers are still more comfortable printing out relevant documents and making notes in margins. Additionally, researchers will send links for electronic documents to their collaborators and each will individually print and make margin annotations. It is not uncommon for intelligence analysts, a specific type of knowledge worker that the authors have experience working with, to spend significantly more of their time collecting material in their analysis. In this work, we aim to address both the quantity of data problem as well as making use of electronic media to increase collaboration and productivity. We do this through a collaborative wiki environment designed to find and filter input data, allow for user input and annotations, and provide a workspace for team members. This system is also designed to link data from sources directly to a research area for maximum productivity and pedigree. In this manner, we're hoping to utilize an approach to even out collection and analysis time and effort to a more reasonable ratio.

System Concept & Design

The fundamental concept behind KEF is of an environment that can act as an assistant to a research team. By providing some documents (e.g., research articles from a domain of focus) as an indication of interest, elements of the KEF environment can automatically identify new and potentially related material, inserting this back into the environment for review (Chappell, 2007). KEF can be configured to harvest information from individual sites, use search engines as proxies, or collect material from social media sites such as blogs, wikis, and forums etc. Harvesting strategies include:

- simple metadata extraction (e.g., author and co-author, material source (e.g., journal name), citations within original documents, etc)
- topic identification (e.g., climate-change, food supply, access to education, etc)
- sentiment analysis (e.g., the fact that the statements related to climate-change are positive or negative)
- rhetorical analysis (e.g., identification of issues being relayed from a protagonist to a target audience with a specific intent to cause an effect) (Sanfilippo, et al., 2008).

Initial results may lack close relevance due to the general criteria for search. Users can vet the material collected, either by single items or by groups (e.g., everything from a particular author or journal). This procedure serves as input to the harvesting strategy until a tightly defined harvesting strategy matches exactly with what the research team needs. Eventually, the research team can expect to receive a steady stream of relevant traditional material and social media.

As the data repository is populated with relevant material, users can interact with the data on a variety of levels depending on their goals. All data in the repository is automatically tagged with basic document metadata (source, author, date, etc.), as well as with semantic information extracted from the text during the ingestion routine. Using information extraction tools, all entities (people, locations, events, etc.) in the text are marked and user-identified key terms are automatically tagged (e.g., climate terms in the case of a climate modeling scenario). These tags provide a means of search and organization that provide for ease of recall. Importantly, users can correct existing annotations, or create their own to match their individual needs. Users can replace manual margin mark-up with notes or annotations that can be searched on later or used by other collaborators. Finally, each document has a talk page where users can discuss (asynchronously) the document (a synchronous 'chat' component is also available).

The Process

From a users perspective, the KEF process is illustrated in Figure 1.

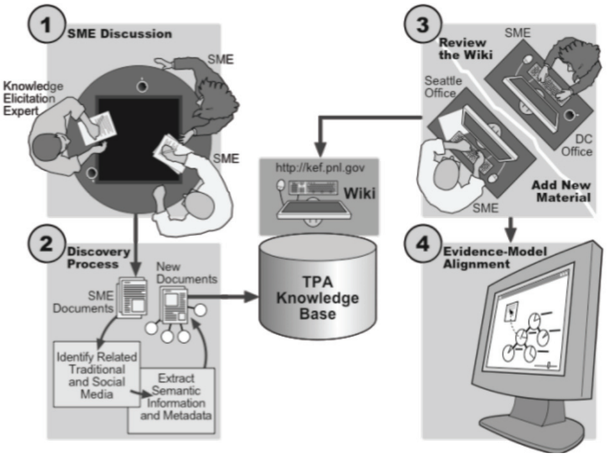
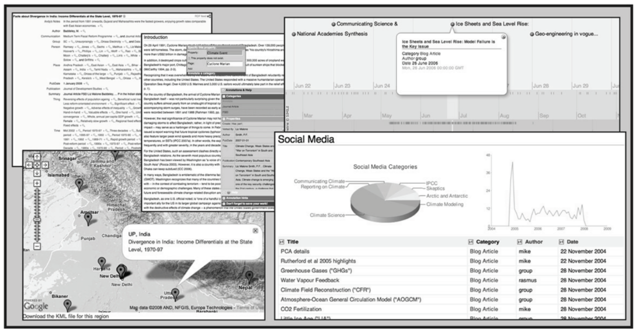


Figure 1. The KEF Process

Knowledge elicitation experts meet with modelers and subject-matter experts[1] to get an understanding of their problem. For example, in the case of a modeling group trying to understand the effects of climate change on the Indian sub-continent, this may lead to the creation of a context map showing all the elements of climate change that may apply (e.g., access to education, clean water, etc) and a selection of documents currently used to create and parameterize their models.

Documents collected in this first phase are used as part of the discovery phase. The documents are 'virtually' dissected by a number of KEF components (i.e., automated software tools) in order to understand their constituents and relevance. Based on these elements, new material (e.g., documents, websites, blogs, forums, news articles, etc) are discovered and pushed through an extraction pipeline prior to being ingested into the knowledge base. This process is cyclic, altered by the feedback provided by the user during the vetting/review phase.

As material is introduced to the knowledge base, it can be reviewed by the users through the KEF wiki (Figure 2), through a number of content-specific views utilizing graphs, maps, etc. The wiki provides a simple but powerful collaborative environment for the vetting, evaluation and alignment of evidence to models.



Ice Sheets and Sea Level Rise: Model Failure is the Key Issue
Category Blog Article
Author group
Date 26 June 2006
Mon, 26 Jun 2006 00:00:00 GMT

UP, India
Divergence in India: Income Differentials at the State Level, 1970-97

Social Media

Title	Category	Author	Date
PCA details	Blog Article	mike	22 November 2004
Rutherford et al 2005 highlights	Blog Article	mike	22 November 2004
Greenhouse Gases ("GHGs")	Blog Article	group	28 November 2004
Water Vapour Feedback	Blog Article	rasmus	28 November 2004
Climate Field Reconstruction ("CFR")	Blog Article	group	28 November 2004
Atmosphere-Ocean General Circulation Model ("AOGCM")	Blog Article	group	28 November 2004
CO2 Fertilization	Blog Article	mike	28 November 2004
Little Ice Age ("LIA")	Blog Article	group	28 November 2004

Figure 2. Multiple Views.

Conclusion

We have presented a collaborative workspace for researchers to gather, annotate, share and store relevant information. The combination of automatically harvested and annotated material with user vetting helps the researcher effectively handle the potentially large quantities of data available, while providing a measure of quality control.

References

Sanfilippo, A.P., Franklin, L., Tratz, S.C., Danielson, G.R., Mileson, N.D., Riensche, R.M. and McGrath, L. 2008. Automating Frame Analysis. *In Social Computing, Behavioral Modeling, and Prediction*. eds. Liu, H.,. Salerno, J.J, and Young, M.J. New York, NY : Springer. 239-248.

Chappell A.R., Posse, C., Willse, A.R., Donaldson, A. and Tratz, S.C. 2007. Concept-Based Document Analysis. [Unpublished]

[1] Depending on the domain, these may be the same person.