Semantic Information Representation and Ontologies

1. Research Team

Project Leader:	Prof. Dennis McLeod, *Computer Science*
Other Faculty:	Prof. Isaac Cohen, *Computer Science* Prof. Shrikanth Narayanan, *Electrical Engineering* Prof. Cyrus Shahabi, *Computer Science*
Graduate Students:	Anne Yun-An Chen, Seokkyung Chung, Vesile Evrim, Shan Gao, Vijayakumar Gopalakrishnan, Jong-eun Jun, Leila Kaghazian, Hyun Woong Shin, Bomi Song, Sang-Soo Sung, Seongwook Youn
Industrial Partner(s):	JPL

2. Statement of Project Goals

The primary goal of this project is to design, develop, deploy, and test methodologies for exploiting semantic aspects of the data. Toward this end, we have developed an information system based on ontologies (collections of key concepts and terms along with their inter-relationships) to provide intelligent information selection. The main capabilities of the system include modeling the meaning of the data, and performing mining operations from information streams.

A second aspect of this project has been focused on intelligent information presentation by utilizing abstractions of narrative story, which is referred to as story models. Story models define how content should be combined and formed. By selecting the best story model for a user's request, we can deliver a customized story by combining information content that matched their question according to the specifications of story models. Since story models can include multimedia data as well as text articles, exploiting story models will enable creation of dynamic multimedia presentations.

3. Project Role in Support of IMSC Strategic Plan

Our research has a strong connection with IMSC three Vision Projects. With regard to communication vision project (whose goal is to create an immersive environment for trainees and experts to minimize social presence), our main role is to implement the local agent that helps the trainee to fix the computer by displaying the multimedia information. In addition, we produce the state-summary report that lets experts know the history of the trainee's behavior when he or she calls the experts. To fulfill these requirements, we will build a Task Domain Ontology, and state-summary reports using story models.

4. Discussion of Methodology Used

The most critical issue in intelligent information management is how to represent and extract the semantic meaning from information contents. This problem has been addressed in diverse research areas including artificial intelligence, information retrieval, natural language processing, multimedia, etc.

In this project, we choose an approach based on a domain-dependent ontology to express semantic knowledge. In particular, we seek dynamic ontology management since ontologies must evolve with time as the concepts in that domain change. Our research approach can be also referred to as multi-disciplinary in that we use the ideas from diverse areas including data mining, information retrieval, machine learning, semantic web, natural language processing, signal processing, etc.

5. Short Description of Achievements in Previous Years

Our previous research efforts spread across three areas: information presentation based on story models, topic mining from news streams, and development of Task Domain Ontology for the communication vision project. Although these three areas seems to be independent research efforts at first glance, they are inter-connected each other through several IMSC projects, including the I4 (Immersive, Interactive, Individualized Information), and communication effort.

I4 is aimed to immerse people in interactive, customized multimedia information experiences and to provide tools for maintaining the underlying content system. Rather than issuing a query to an information system and getting back some search results, a user will receive an integrated information presentation or "story". To this end, we created domain-dependent ontology, and a library of multi-modal content that is related through a rich ontology. From complete text stories to video clips, all information content is tied into a common ontology that is used to extract elements for stories. The story models consist of a small series of manually constructed templates that leverage hand-authored content. User requests are processed and matched to a particular story model, and all content that matches the request is flown into the model to create an original story. Thus, a story model is a central concept in I4. In addition, the capability of topic mining in identifying new patterns (e.g., events, concepts, key terms) from news streams can be coupled with dynamic ontology management, which is also essential in I4.

With regard to Communication Vision Project, we developed a story model that defines several story types that lay out the appropriate presentation style depending on the each main step or exception to fix a computer. Different kinds of story types will have different intentions and goals. A trainee will receive customized stories in response to their situation. Each story type is filled with combinations of information elements such as video, audio, images and texts. Moreover, story types employ visual techniques that solve layout problems such as the way different kinds of information are combined and presented effectively. This approach has almost same idea with domain-independent story model introduced in [14].

5a. Detail of Accomplishments During the Past Year

Information Presentation Based on Story Models

In today's Web service industry, information presentations and collections of data are in static format, and they are limited in terms of multimodal presentations. That is, there is very limited capability to dynamically adapt an integrated presentation of information to a user. However, the user will engage deeply into a story when the user, not only reads text articles but also watches videos and/or listens to audio clips in a coordinated manner. In addition, most Web services do not instantiate and customized (for an individual user) "generic" stories.

To address the above issues, we have developed story structures that can be dynamically instantiated for different user requests from various multimodal elements [14]. Furthermore, the system leverages information so that a user will read an appropriate level of story depending upon the user's intention level ranging from general to specific. For example, a user might impress the USC football game, but the user has very little knowledge of the USC football team. The user then wants to read very general information instead of specific information regarding the team. When the user requests a general level of the USC football team, the system will deliver a customized (in this case, a general story) dynamically generated multimodal story.

In order to convey the nature of the information presentation, we have developed the precise nature of the dynamic generation of user-customized multimedia presentations that will draw upon visual techniques, presentation constrains, a content query formulation, a story assembly and a structured rule-based decision process.

The proposed story model defines four story types that lay out the appropriate presentation style depending on the user's intention and goal – a summary story type, a text-based story type, a non-text based story type and a structured collection story type. Different kinds of story types will have different intentions and goals. Consumers will receive customized stories in response to their requests. Moreover, we design a story model to delineate high-level abstractions of general story templates so that the proposed story types can cope with any kind of existing stories.

The idea of a story type is generic in that it can be "filled in" with combinations of information elements (objects). Moreover, story types employ visual techniques that solve layout problems such as the way different kinds of information are combined and presented effectively. The key components of the story type include the following:

- *Title*. It specifies the nature of presentation.
- *Elements*. Information objects from the Content Database that are required and can fit into an instantiation of the story type.
- *Element description*. It defines elements to be instantiated such as a text-image style or a text-image-audio-video style.

To determine a user's intention and goal, a general knowledge-based process, with selection (information filtering) heuristics, is used. A key to the successful use of story types is the ability to relate and connect the user requests to the content database. A domain dependent ontology is essential for capturing the key concepts and relationships in an application domain; metadata descriptions will connect a modified user request (by using domain ontology) to the content database for retrieving proper content elements.

Metadata descriptions are annotations of content objects that describe high-level information. They describe the content, quality, condition and other characteristics of multimedia objects. We design two types of annotations to delineate content objects. The first type of annotation depicts associations between concepts and content objects, a level of generality spectrum and a level of relevance spectrum. This type of annotation will be used to retrieve proper content objects dependent on a modified user's request. The levels of generality spectrum are the core part for information leverage. This generality spectrum is divided into 10 levels from general to specific. The other type of annotation denotes content objects themselves – titles, descriptions, media types, and locations. This type of metadata will be used to fill out a determined story type. These meta-descriptions are critical to instantiate appropriate content objects for a story type that are determined by a structured rule-based decision process.

Topic Mining from News Streams

As the Web continues to grow as a vehicle for the distribution of information, many news organizations are providing newswire services through the Internet. Given this popularity of the Web news services, it is necessary to provide automatic data mining tools that allow the users to quickly meet their information needs. This facilitates information navigation and search, and at the same time presents an efficient framework for managing a document repository, as the number of documents grows extremely huge.

The simplest document access method within Web news services is keyword-based retrieval. Although this method seems effective, there exist at least three serious drawbacks. First, if a user chooses irrelevant keywords (due to broad and vague information need or unfamiliarity with the domain of interest), retrieval accuracy will be degraded. Second, because keyword-based retrieval relies on the syntactic properties of information (e.g., keyword counting), a semantic gap cannot be overcome. Third, only expected information can be retrieved because the specified keywords are generated from users' knowledge space. Hence, if the users are unaware of the airplane crash that occurred yesterday, then they cannot issue a query about that accident though they might be interested.

The first two drawbacks stated above have been addressed by query expansion based on domain-independent ontologies. However, it is well known that this approach leads to a degradation of precision. That is, given that the words introduced by term expansion may have more than one meaning, using additional terms can improve recall, but decrease precision. Manually building an ontology with a controlled vocabulary is helpful in this situation. However, although ontology-authoring tools have been developed in the past decades, manually constructing ontologies whenever new domains are encountered is an error-prone and time-consuming process. Therefore, integration of knowledge acquisition with data mining, which is referred to as ontology learning, is necessary [12].

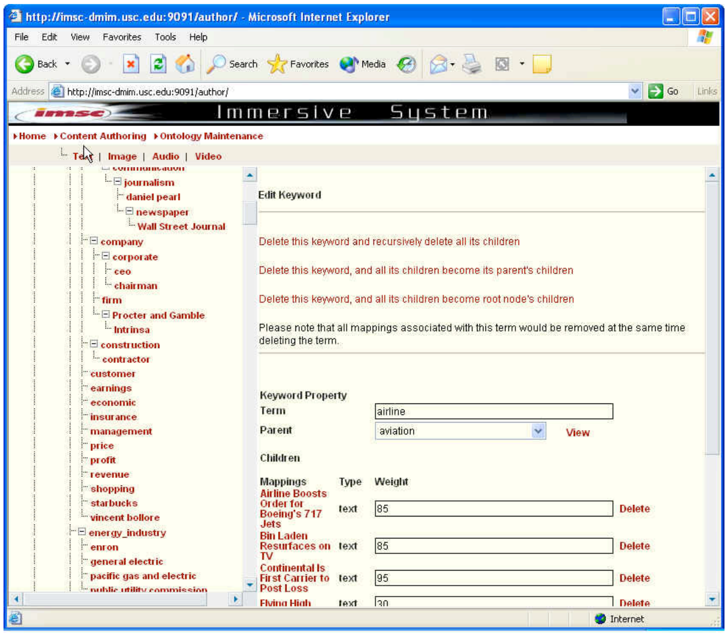
To address the above three drawbacks, we have developed *topic mining*, which effectively identifies useful patterns (e.g., metadata, topics, events) from news streams [2]. News articles are retrieved from Web news services on a daily basis, and processed by data mining tools to produce useful higher-level knowledge, which is stored in a content description database. Instead of interacting with a Web news service directly, by exploiting the knowledge in the database, an information delivery agent can present an answer in response to a user request. Current capabilities on topic mining from news stream datasets include the following:

- *Efficient incremental hierarchical news document clustering*. Since several hundred news stories are published everyday at a single Web news site, to cope with such dynamic environments, we should provide efficient incremental data mining algorithms. Despite the huge body of research efforts on document clustering, little work has been conducted in the context of incremental hierarchical news document clustering. Our developed clustering algorithm based on a neighborhood-search has several key advantages, including the scalability with the high dimensionality, capability to discover clusters with different shapes and sizes, and ability to provide succinct description of clusters.
- *Topic detection and tracking*. Due to the overwhelming amount of information involved, it is crucial to provide an intelligent agent that can identify novel information and track related information for a user. Given a stream of news articles**,** topic mining identifies whether a new document belongs to an existing topic or new topic. Topic mining also tracks events of interest based on sample news story. For example, it associates incoming news stories with the related stories (which were already discussed before), or it can also monitor the news stream for further stories on the same topic.
- *Topic ontology learning from a news stream*. In order to achieve rich semantic information retrieval, metadata (e.g., ontological information) should be employed. Since manually building and maintaining such metadata is nearly impossible, we developed a prototype system for learning topic ontologies. A topic ontology is a collection of concepts and relations. One view of a concept is as a set of terms that characterize a topic. We employ two generic kinds of relations, specialization and generalization. The former is useful when refining a query while the latter can be used when we generalize the query to increase recall or broaden the search.

An experimental prototype system has been developed, implemented and tested to demonstrate the effectiveness of the topic mining framework. The results show that the proposed clustering algorithm produces high-quality document cluster hierarchy, and obtained topic ontology provides an interpretation of the news topics at different levels of abstraction. In addition, we also developed the Web-based ontology authoring tool as shown in Figure 1. This authoring tool can be extended into topic ontology editing.

Development of task domain ontology for Communication Vision Project

Our main focus in communication vision project is to build a local task assistance agent. As illustrated in Figure 2, all behavior of the trainee is observed and the information is stored back into “User Sensing Profile”. When the trainee requests “What’s next?”, then the “Task Plan Representation (Planning)” determines the specific next task by using “Task Domain Ontology” and “User Sensing Profile”. The determined task will be sent to the “Task Plan Actions”, and we generate a proper presentation according to the determined task. Moreover, the local agent can generate summary-reports for the experts.



Edit Keyword

Delete this keyword and recursively delete all its children

Delete this keyword, and all its children become its parent's children

Delete this keyword, and all its children become root node's children

Please note that all mappings associated with this term would be removed at the same time deleting the term.

Keyword Property

Term: airline

Parent: aviation View

Children

Mappings	Type	Weight	
Airline Boosts Order for Boeing's 717 Jets	text	85	Delete
Bin Laden Resurfaces on TV	text	85	Delete
Continental Is First Carrier to Post Loss	text	95	Delete
Flying High	text	30	Delete

Figure 1. Sample screen-shot for Web-based ontology-authoring tool on news domain

Thus, Task Domain Ontology plays a key role in building a local agent. The Task Domain Ontology formally specifies terminology associated with the type of tasks, its exceptions and the status of the exception. Node selection is based on not only user requests but also user sensing data in planning module. Therefore, we need planning that is a well-defined logical model for making decision and plans to retrieve proper information from the Task Domain Ontology. The Task Domain Ontology is obtained through each step of scenario, as well as information provided by human factor experts. Each step contains several sub-cases that represent either success case or exception cases. In addition, a sub-case is determined by preliminary results that observed by several human tests.

6. Other Relevant Work Being Conducted and How this Project is Different

In the past few decades, representing metadata has been investigated in diverse disciplines including programming languages, database management, and artificial intelligence. Recently, XML is proposed as a standard for representing and exchanging data (and metadata). However, XML can only reflect syntactic property of data, and it has a limitation in representing semantic aspect of information. The Semantic Web research efforts, including developing data model such as RDF, DAML+OIL, and OWL [3.4.6], are closely related with our project. We will complement our work by exploring and implementing existing models in these fields.

With regard to uniqueness, topic mining for ontology management can be distinguished from other research in that we are investigating ontology learning algorithms on information streams, rather than static dataset.

7. Plan for the Next Year

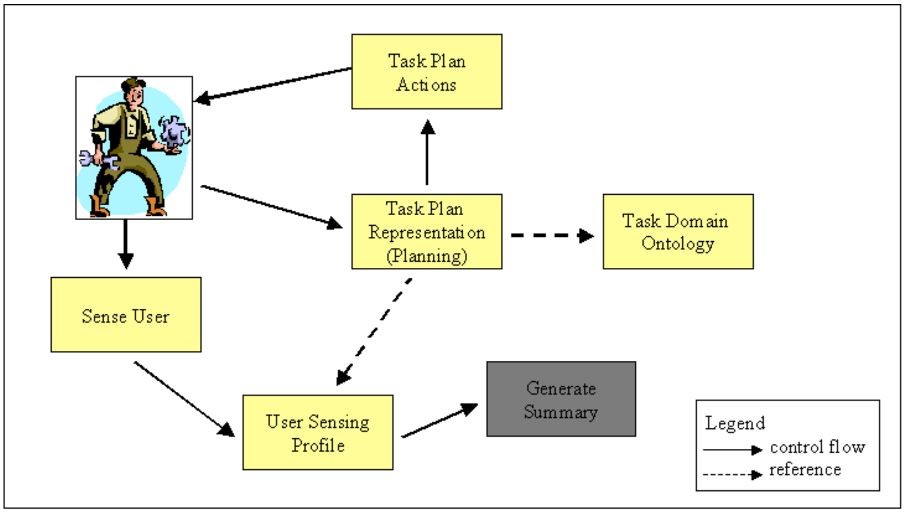


Figure 2. Illustration of a local agent

Leading our plans for next year is a careful analysis of the use of our research outcomes in three Vision Projects: education, communication, and entertainment. Our initial main focus will be how to efficiently apply task ontology in Communication Vision Project.

Another key focus will be the applicability of topic mining to ontology management system to accommodate rich semantic information extraction. To this end, we will annotate topic ontology within Protégé WordNet tab [13] for news applications.

Moreover, to strengthen our work in generality of application perspective, we will extend the topic mining framework into other datasets, e.g., intelligence data. Since data in intelligence applications (e.g., homeland security) arrives as high-speed information streams, we need to process this decision-critical information very quickly. Thus, our incremental data mining module is expected to play a key role in processing incoming data online.

8. Expected Milestones and Deliverables

Spread across the four-year scope of this project, we will provide several deliverables including topic mining modules for intelligence applications, and topic ontology implementation within an ontology management system. To accomplish the deliverables, each milestone represents the culmination of research and the realization of the project requirements.

Annotate topic ontology using Protégé WordNet tab (Year 2):
The WordNet tab allows users to search WordNet and annotate Protégé ontology with terms, synonyms, and relations from WordNet. That is, it allows the users to create new Protégé classes and instances with WordNet content. Thus, to obtain rich semantics for topic ontology, coupling with WordNet, we will implement topic ontology within Protégé.

(Semi-) Automated ontology augmentation (Year 2-3):
This milestone represents an important aspect of the system – the ability to create new nodes in the underlying ontology with little user intervention. When unknown (strong) concepts are encountered in processed content, the system will attempt to create new nodes in the ontology for those concepts. This represents important information management research into ontology construction. The developed topic mining module and ontology management system will minimize human efforts in this process.

Use of advanced/rich story models (Years 2-3):
The full vision of the story models requirement will be realized. Story models with rich specifications can be used with any content, regardless of its origin, without requiring manual authoring of content. The system will perform any conversion or editing of content in order for it to be placed within a particular model.

GUI creation and editing of story models and package (Years 2-3):
An organic authoring interface for story models will allow editors to initially create story packages through a graphic user interface. Editors will be able to drag stories together to create new structures for displaying related content. Eventually, using a non-technical, graphic user interface, editors can create full story models with rich specifications.

Refinement of task domain ontology (Year 2-3)
Currently, the planning is based upon a structured rule-based formalism. This representation formalism may evolve into logic-based formalism. This allows a robust mathematical framework where all concepts can have complete and formal expression [11].

Development of task assistance state-summary report (Year 2-3)
In the Communication Vision Project, the state-summary report consists of key multimedia data from user sensing profile. This state-summary report allows the experts to understand what the trainee did before they connected. In addition, the experts send the reports to other experts when they need advices. As a result, experts can understand the situation as if they watch the whole behaviors of the trainee in front of them.

Applicability of topic mining to intelligence applications (Year 2-4):
We plan to extend topic mining to other applications to show our framework can be generalized and utilized for others. Specifically, we decided to apply our methodology to intelligence applications for homeland security.

In homeland security, since data arrives via high-speed information streams, we need to process this decision-critical information very quickly. Hence, the mining algorithm must be able to process this information stream in one-pass, and leverage the data in real-time. That is, it should be equipped with a capability of mining data online, and identifying (anomalous) patterns continuously. This online mining ability is particularly important in homeland security applications since alerting warning about potential threats as early as we can is critical. Furthermore, intelligent data can be distributed over several law enforcement organizations (e.g., FBI, CIA). These data sources can be heterogeneous (i.e., different database systems, different database schemas, etc) since they are usually maintained by different agencies. Thus, sharing and exchanging the information effectively becomes a must. In sum, monitoring information streams and sharing/exchanging heterogeneous information are two functions in homeland security, upon which we plan to focus.

Our incremental clustering module in topic mining can process incoming data online, and effectively identify useful patterns. In addition, we plan to develop a novelty detection module, which identifies potential anomalous events for the purpose of intrusion detection or crisis alert. Moreover, our research efforts on news dataset mining [2.7.8.9.10] can be directly applicable to analyzing vast amount of unstructured data (e.g., emails, news streams, field reports, web pages). Finally, our research accomplishments on database federation and semantic metadata management [1.5] can effectively deal with many issues on sharing intelligence data, e.g., conceptual and schematic diversity, information sharing and exchange, knowledge discovery, etc. Therefore, it is worthwhile to extend topic mining into homeland security since our methodologies are technically feasible and directly appropriate.

9. Member Company Benefits

We have a new direct connection with earthquake science research at JPL, which has strong requirements for the kinds of ontology and user modeling techniques we are developing. We now have substantial support from NASA on an application study directly connecting to the research (see http://www-aig.jpl.nasa.gov/public/dus/quakesim/index.html for the description of QuakeSim project).

10. References

[1] A.Y. Chen, A. Donnellan, D. McLeod, G. Fox, J. Parker, J. Rundle, L. Grant, M. Pierce, M. Gould, S. Chung, and S. Gao. Interoperability and semantics for heterogeneous earthquake science data. In *Proceedings of the International Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*, 2003.

[2] S. Chung, and D. McLeod. Dynamic topic mining from news stream data. In *Proceedings of ODBASE*, 2003.

[3] S. Decker, D. Brickley, J. Saarela, and J. Angele. A query and inference service for RDF. In *Proceedings of the Query Language Workshop*, 1998.

[4] D. Fensel, F. van Harmelen, I. Horrocks, D. McGuiness, and P. Patel-Schneider. OIL*:* An ontology infrastructure for the Semantic Web, *IEEE Intelligent Systems*, 16(2): 38-45, 2001.

[5] M. Hammer, and D. McLeod. Database description with SDM: A semantic database model. *ACM Transactions on Database Systems*, 6(3): 351-386, 1981.

[6] J. Heflin and J. Hendler. A portrait of the Semantic Web in action, intelligent systems. *IEEE Expert*, 16(2): 54 -59, March-April 2001.

[7] L. Khan, and D. McLeod. Audio structuring and personalized retrieval using ontologies. In *Proceedings of IEEE Advances in Digital Libraries, 2000.*

[8] L. Khan, and D. McLeod. Effective retrieval of audio information from annotated text using ontologies In *Proceedings of ACM SIGKDD Workshop on Multimedia Data Mining*, 2000.

[9] L. Khan, and D. McLeod. Disambiguation of annotated text of audio using ontologies. In *Proceeding of ACM SIGKDD Workshop on Text Mining*, 2000.

[10] L. Khan, D. McLeod, and E. Hovy. Retrieval effectiveness of an ontology-based model for information selection. *The VLDB Journal*, 13(1): 71-85, 2004.

[11] S. Krivov, A. Dahiya, and J. Ashraf. From equations to patterns: logic based approach to general systems. International Journal of General Systems, 31(2): 183-205, 2002.

[12] A. Maedche, and S. Staab. Ontology learning for the Semantic Web. *IEEE Intelligent Systems*, 16(2), 2001.

[13] N.F. Noy, M. Sintek, S. Decker, M. Crubézy, R.W. Fergerson, and M.A. Musen. Creating Semantic Web contents with Protégé-2000. *IEEE Intelligent Systems*, 6(12): 60-71.

[14] H. Shin, D. McLeod, and L. Pryor. The dynamic generation of user-customized multimedia presentations. *Accepted for publication in IRMA*, 2004.

Note: Further information on our research can be found on our website at http://imsc-dmim.usc.edu.