Knowledge on Demand:
Human Language Technology for Knowledge and Expertise Discovery

Mark Maybury

The MITRE Corporation
202 Burlington Road
Bedford, Massachusetts 01730
maybury@mitre.org
http://www.mitre.org/resources/centers/it

Abstract

For several years we have been pursuing a vision of knowledge on demand, the ability for all users to access knowledge regardless of time, location, device or level of expertise. This paper overviews the central role human language technology can play in several areas of knowledge on demand. After motivating the importance of this area, we describe several implemented examples of human language technology applied to knowledge management functions including discovery of knowledge and discovery of experts.

Knowledge on Demand

Our vision for Knowledge on Demand is the ability of all users to access tailored knowledge sources and services regardless of time, location, interactive device or level of expertise. As shown in Figure 1, this fits within a larger context and process of knowledge management. The process of managing knowledge includes, but is not limited to, the demand for, creation or discovery of, and delivery of pre-existing or new knowledge. Important enabling tasks include discovery of explicit knowledge (e.g., in databases, in corporate documents, on the web), discovery of experts, formation of expert teams and/or communities, creation of new knowledge, and tailoring delivery of knowledge to the recipient.

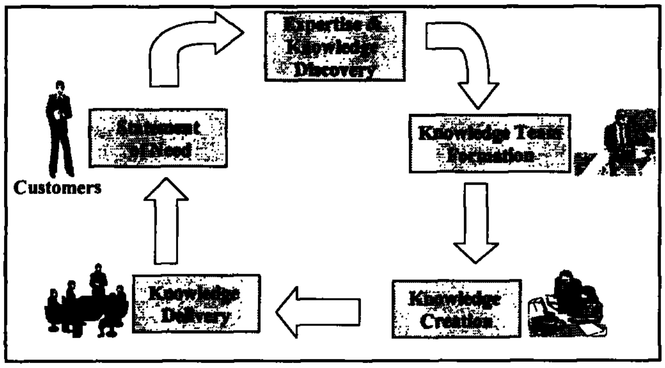


Figure 1. Knowledge Management

Human Language Technology

Human language technology (HLT) promises significant capabilities that can advance knowledge management solutions in the face of growing volumes of global knowledge and the acceleration of the pace of knowledge change. Consider that in the US alone there are more than 300,000 new patent applications annually which result in approximately 160,000 new patents added to the more than 6 million current patents. This volume of knowledge and knowledge artifacts makes Al Roger's statement all the more pertinent: "In times of profound change, learners inherit the Earth, while the learned find themselves beautifully equipped to deal with a world that no longer exists."

Whereas much knowledge is implicit in experts' minds, frequently they express their competence either in written or verbal form or via their utilization of computer programs or the web. Consider that the size of the Library of Congress is 33 terabytes (growing at about 7,000 materials a day), whereas one estimate is that the long distance communications in the U.S. alone in 1999 were 70,000 terabytes. Patents, libraries, and conversations are examples of mechanisms that capture and express knowledge in spoken and natural language, knowledge that is essential to competitive advantage in the global marketplace.

Human language technology promises capabilities to facilitate access to such onerous collections by providing facilities that:

- create indices of corporate knowledge using methods such as information extraction, clustering, and visualization
- summarize large volumes of knowledge, including those produced and accessed by experts

- transcribe audio or video sources which often capture expert communications and/or lessons learned
- translate foreign language material enabling access to global knowledge sources
- generate profiles of experts using language processing to enable automated expert finding
- facilitate communities of interest by classifying and tracking public user and group interests and sharing these to enhance group awareness

We have developed and deployed a range of applications that take advantage of human language. This paper focuses on the challenge of global and corporate knowledge management and outlines experience with intelligent tools that support:

- automated detection and tracking of emerging topics from unstructured multimedia data
- automated discovery of distributed experts and communities of expertise, and
- capabilities to increase organizational awareness (e.g., awareness of team members and materials in virtual collaboration environments).

To illustrate these areas concretely, we overview several implemented and evaluated systems. Figure 2 illustrates how these applications are envisioned to work together to facilitate knowledge on demand. Overall, the systems work together to enable a user to retrieve multimedia documents, extract information from those, summarize their contents, translate them if they are in a foreign language, cluster and mine related collections of documents, or browse within a geospatial context collections of documents or information extracted from those documents (e.g., topics, names, locations). In addition, we have a need to find experts and facilitate their collaboration and enable new knowledge creation.

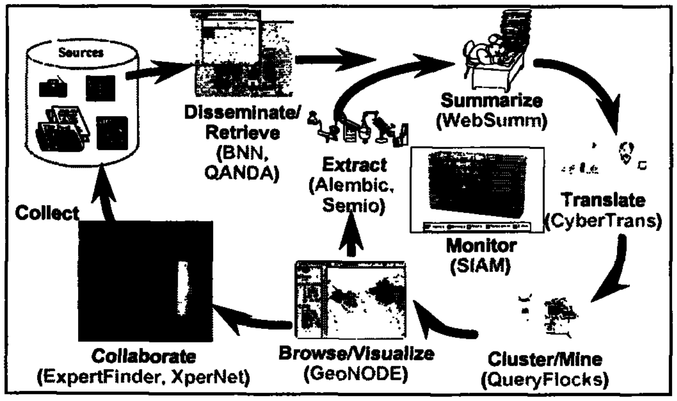


Figure 2. Human Language Technologies for Knowledge Management

In the remainder of the paper we describe several of the systems in Figure 2 including Geospatial News On Demand Environment (GeoNODE) (Hyland et al. 1999), Expert Finder, XperNet and the Collaborative Virtual Workspace (CVW). GeoNODE automatically detects and extracts emerging topics from web, document, and video news broadcast collections using a unique underlying data mining algorithm (QueryFlocks). The Broadcast News Navigator (BNN) provides content-based access to news for GeoNODE. The Semio system, which we do not describe, similarly provides automated extraction and visualization from unstructured data, to include browsing by topic in a Yahoo-like fashion across ingested collections. QANDA is a question answering system (Breck et al. 2000) and WebSumm summarizes documents. CyberTrans provides document translation on demand using a range of commercial translation engines. Expert Finder is a people skill finder that exploits the intellectual products created within an enterprise to support automated expertise classification. XperNet addresses the problem of detecting extant or emerging communities of human expertise without a priori knowledge of their existence. Both ExpertFinder and XperNet combine to detect and track experts and expert communities within a complex work environment. CVW (cvw.soureforge.net) is a place-based collaboration environment that enables team members to find one another and work together. These are described in turn.

Knowledge Discovery

Given rapid knowledge creation and dissemination, a key technical challenge is dealing with large, heterogeneous collections of knowledge in a uniform manner. We have been exploring a range of analytic support tools to facilitate users in discovering and accessing knowledge on demand. Because these tools are fully automated, we are limited to discovering knowledge that is explicit in artifacts and can be extracted, to include information about people, places, organizations, and events. For example, Figure 3 illustrates the architecture of MITRE's GeoNODE. GeoNODE aims to provide uniform access to unstructured multimedia (i.e., text, audio, and video) from such sources as document repositories, broadcast news (e.g., CNN, MS-NBC), and the World Wide Web. Following initial data preprocessing (possibly including speech transcription) into a common format (e.g., to delineate segments of stories, web pages, or messages), multilingual information extractors tag named entities (e.g., names of people, organizations, and locations) as well as relations among these (Aberdeen et al. 1995). These named entities are then clustered according to co-occurrence within segments and then a graph analysis routine is used to further segment these clusters into "topics". The most frequently occurring named entities within each topic cluster become the topic label for that set of entities, which in turn point back to source segments.

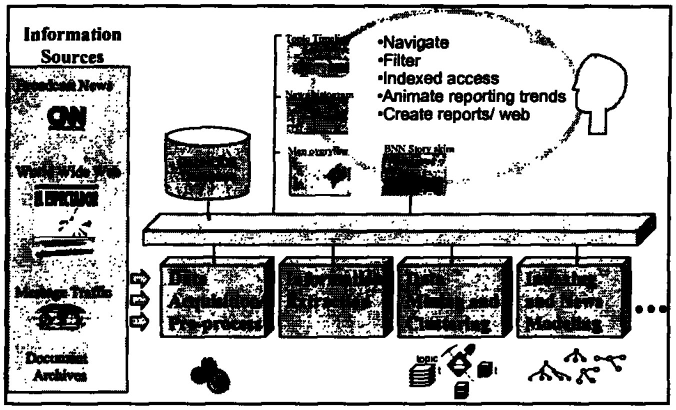


Figure 3. Geospatial News on Demand Architecture

This unique pipeline results in uniform direct access/visualization across the heterogeneous sources. As Figure 3 suggests, this enables a range of customized views including the ability to visualize news events temporally as well as geospatially across sources. Using this heterogeneous access, an information analyst can quickly assess key individuals and events reported from a range of publications across sources. For example, this would enable a financial analyst to track local business and government leaders, corporations, and key economic events to assess the financial state of a region.

Figure 4 illustrates a simple visualization over a six month period of the occurrences of reporting on a topic (i.e., the number of detected stories) across two broadcast news sources (CNN World View and CNN World Today). Interactive tools allow the user to quickly zoom in on relevant time periods to focus on unusual or interesting levels of reporting on particular detected topics.

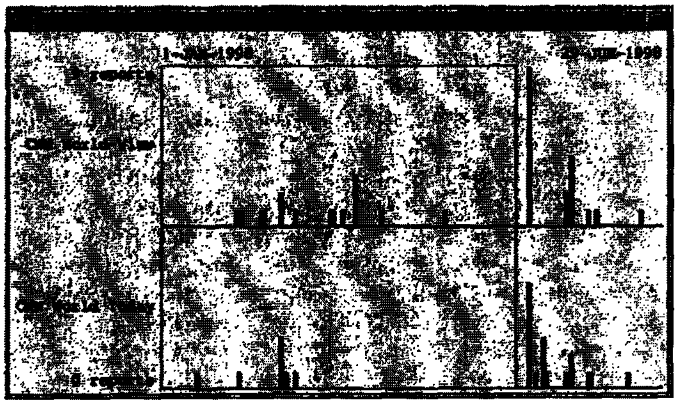


Figure 4. Visualizing Relevant Stories by Source

A more sophisticated visualization displays topic occurrences over time. In the visualization in Figure 5, the user has selected the two-year timeframe from October 1996 until November of 1998. GeoNODE displays automatically discovered topics (list on bottom right of Figure 5) that are reported across particular news programs (bottom left hand list). Notice how topics discovered by GeoNODE are labeled using the most frequently occurring named entities in the clusters. In this case the topic cluster "Afghanistan, Kenya, Nairobi, Tanzania" refers to documents concerning the US embassy bombings and the topic cluster "Afghanistan, Khartoum, Ladin, Sudan" refers to stories regarding the US air strike response. When then user selects (highlights) topic clusters, GeoNODE displays the intervals of occurrence in the top display of Figure 5. Selection of clusters enables direct access to the underlying documents forming the topic as well as other services such as translation or summarization.

Figure 5. Temporal Visualization of Automatically Nominated Topics

A final form of visualization is a custom generated ArcView Gis cartographic display. Figure 6 illustrates how a user can take a particular set of documents associated with an identified topic and visualize the frequency of mentions of locations in documents on those topics. For example, Figure 6 shows a heavy concentration of reporting on Indian and Afghanistan on the topic of nuclear testing. Shades of brown indicate topics reported about regions (e.g., a country) whereas yellow circles display concentrations of reporting in particular cities.

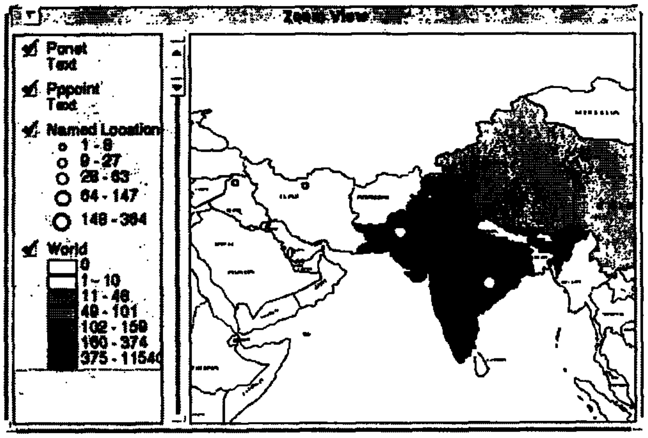
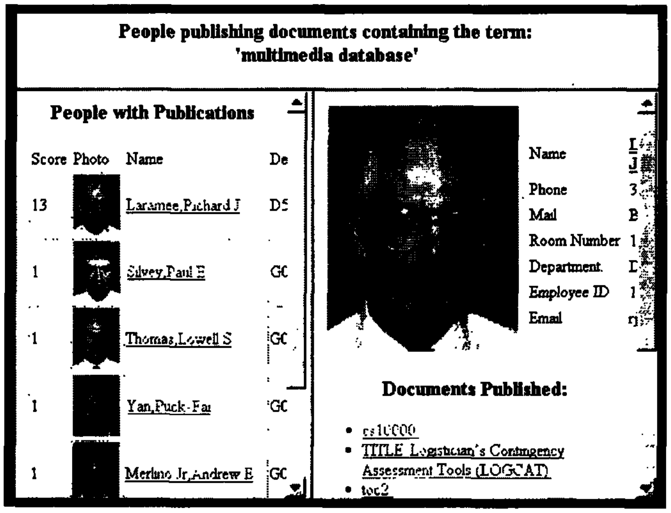


Figure 6. Geospatial Visualization

In summary, temporal, spatial and topical visualizations enable the user direct access to extracted knowledge to support rapid browsing, comparison, anomaly and pattern detection.

Expert Finding and Community Discovery

While access to extracted information and discovered relationships is important, much knowledge is not captured explicitly, but rather implicitly resides in processes, procedures, or experts' heads. Human language technology can facilitate implicit knowledge discovery by extracting information about experts from their writings or in others writings about them. This was precisely the aim of ExpertFinder (Mattox et al. 1999). Given a simple user query, such as "Find me all experts on multimedia databases", ExpertFinder analyzes the most frequently occurring keywords from documents users have published (e.g., resumes, documents, briefings). ExpertFinder also processes corporate newsletters to extract individual names and associated topics that are reported (e.g., a report of an expert presenting a paper at a conference on a particular topic or an interview with an expert). With a database of names associated with topics, the system can now provide expert finding services. Figure 7 illustrates ExpertFinder in action, providing a rank ordered list of "multimedia database" experts based on the frequency and type of evidence the system compiles on their expertise. In Figure 7, the sources of evidence, including links to the original document or data is listed in the bottom right hand corner. A keyword appearing in a resume or in the title of a project for which an individual is a principal investigator is more strongly weighted that the user posting a document containing the keyword. An empirical evaluation of the system (Maybury et al. 2000) in five diverse technical domains found the system to perform approximately one third as well as resource managers.



People publishing documents containing the term: 'multimedia database'

People with Publications

Score	Photo	Name	De
13		Laramee,Pichard J	DS
1		Silvey,Paul E	GC
1		Thomas,Lowell S	GC
1		Yan,Puck-Fai	GC
1		Merlino Jr,Andrew E	GC

Name L J
Phone 3
Mail B
Room Number 1
Department D
Employee ID 1
Email rj

Documents Published:

- cs10000
- TITLE Logistician's Contingency Assessment Tools (LOGCAT)
- toc2

Figure 7. Expert Finder

Human language technology can further enable us to discover not only individual experts but also communities of practicing experts. The XperNet system (Maybury, D'Amore and House, forthcoming) performs its analysis on similar sources as ExpertFinder but clusters users into affinity groups on basis of membership computations. Higher levels of expertise are associated with factors such as document authorship, explicit reference or citation, network centrality, personal Web pages, and project membership. Lower expertise levels reflect fewer expertise indicators and possibly counter-indications such as being a member of the non-technical staff. In an empirical evaluation, XperNet found 70% of manually identified experts (precision) and was 60% correct in its assessment that someone was an expert (recall).

Place-based Collaboration

Part of our vision of providing knowledge on demand includes virtual place-based work places that enable users, regardless of location or time zone, to collaborate synchronously or asynchronously. Figure 8 illustrates the Collaborative Virtual Workplace (cvw.sourceforge.net) which integrates services such as audio/video/text conferencing, data sharing, and navigation within a virtual office building containing other users, documents and tools. As evident even in Figure 8, one major issue is awareness of information sources and users. A range of mechanisms are provided to facilitate awareness of documents (e.g., in the Figure note the labels "Private Data" and "Shared Data"), people (e.g., "Online Users" list, "Users in Room"), and user activities (e.g., "Shared Web Browsing" and "Shared Whiteboard" with user labeled cursors). Even given these facilities, services such as avatars that enable tracking and alerting of events in other rooms are invoked by users to enhance their awareness of people, information, and events. We believe tools such as GeoNODE will enable more effective browsing of knowledge related to entities and events in such environments. Further, ExpertFinder and XperNet support rapid discovery of those who are not virtually present or perceptible, enabling rapid team formation.

Future Research

While there are many disincentives to our vision of knowledge on demand (e.g., scarce expertise, lack of awareness of expertise, validation of expertise, time for knowledge discovery and expert interaction), tools for better accessing and extracting explicit knowledge promise to ameliorate some of these impediments. Many research issues remain open including domain independent event detection and tracking, dynamic creation of expert and expert social models, knowledge delivery tailored to individual expertise, and heterogeneous knowledge integration. We have only begun to explore the possibilities for human language technologies to enable knowledge on demand.

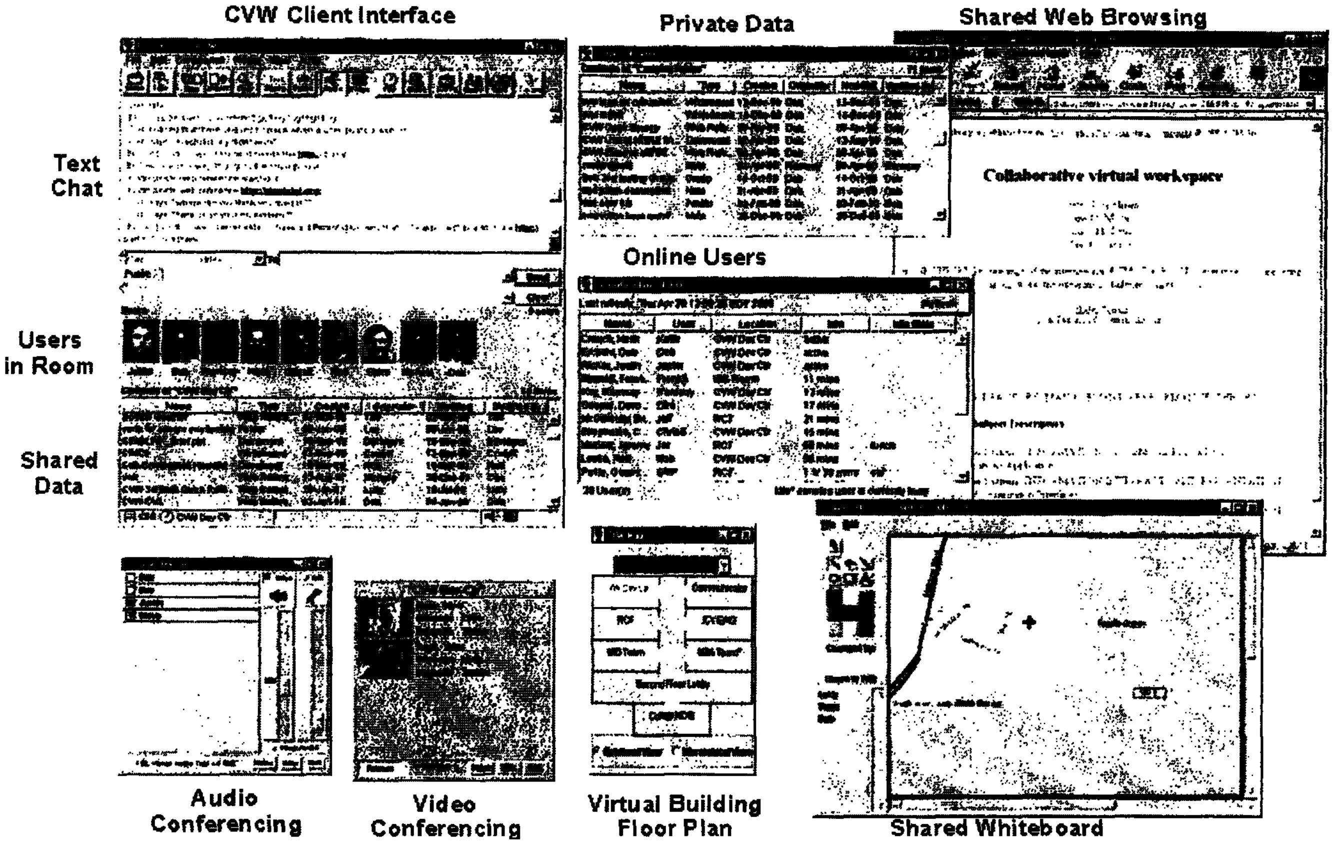


Figure 8. Collaborative Virtual Workplace

References

Aberdeen, A., Burger, J., Day, D., Hirschman, L., Robinson, P. & Vilain, M. 1995. MITRE: Description of the Alembic System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference* (MUC-6), Columbia, MD, 6-8 November 1995, 141-155.

Breck, E., Burger, J. D., Ferro, L., House, D., Light, M., Mani, I. 2000. A Sys called Qanda. In Vorhees, E.. and Harman, D. The Eighth Text Retrieval Conference, NIST Special Pubs. February 2000.

Hyland, R., Clifton, C., and Holland, R. 1999. Geonode: Visualizing News in Geospatial Context. AFCEA Federal Data Mining Symposium. Washington, D.C.

Mattox, D., Maybury, M. and Morey, D. 1999. Enterprise Expert and Knowledge Discovery. International Conference on Human Computer International (HCI 99). 23-27 August 1999. Munich, Germany. 303-307.

Maybury, M., D'Amore, R., House, D. Nov-Dec, 2000. Automating the Finding of Experts. *International Journal of Research Technology Management.* 43(6): 12-15.

Morey, D.; Maybury, M. and Thuraisingham, B. editors, 2001. *Advances in Knowledge Management: Classic and Contemporary Works.* Cambridge: MIT Press.

Maybury, M., D'Amore, R. and House, D. forthcoming. Awareness of Organizational Expertise. *Journal of Human Computer Interaction: Special issue on Awareness.*

Maybury, M. forthcoming 2001. Intelligent Interfaces for Universal Access: Challenges and Promise. In 1st International Conference on Universal Access in HCI (UAHCI), at HCI International, New Orleans, LA, 5-10 August 2001.

Acknowledgements

The work described in this paper represents the efforts of several groups of researchers at The MITRE Corporation including but not limited to John Griffith, Marc Vilain, Dee Goepel, Deb Ercolini, Jay Carlson, Rod Holland, Ray D'Amore, Manu Kanchady, Lynette Hirschman, David Day, Alex Yeh, Ray D'Amore, David House, Inderjeet Mani, Chris Elsaesser, Marc Light, Eric Breck, Stanley Boykin, Warren Greif, Chaya Rosen, Cynthia Small and Jean Tatalias.