[Word Native Format] [RTF Format]

Tools for the Knowledge Analyst:
An Information Superiority Visionary Demonstration

Mark T. Maybury

Advanced Information Systems Center

The MITRE Corporation

202 Burlington Road

Bedford, MA 01730-1420 USA

maybury@mitre.org

http://www.mitre.org/resources/centers/advanced_info/

Abstract

This paper describes the application of a suite of innovative information processing and analytic support tools developed at The MITRE Corporation, in some cases as extensions of commercial tools, to the problem of discovery of illegal chemical proliferation activities from open (i.e. unclassified) sources. These include tools for multimedia information extraction and visualization (BNN), machine translation (WebMT), multidocument summarization (WebSumm), group knowledge management (KEAN), expertise discovery (PeopleFinder), and multiparty collaboration (J-CVW). We consider additional, identified and unfulfilled requirements that suggest future research directions such as data and knowledge mining, intelligent human computer interaction, and semantic collaboration.

1. Introduction

If information is processed data, then knowledge is the creation, evolution, and use of models culled from experience that support understanding, diagnosis, prediction, and other higher level cognitive processes. Increasing volumes of data and information are overwhelming analysts and decision makers and interfere with these higher level and valuable knowledge operations. This paper presents several tools that, in an integrated fashion, are shown to facilitate such higher level reasoning in a user. We present these tools in context of solving a hypothetical but realistic scenario: the discovery of Western companies suspected of supplying chemicals to Iraq for possible use in chemical warfare. Whereas this problem is hypothetical, the scenario utilizes actual data and operational tools.
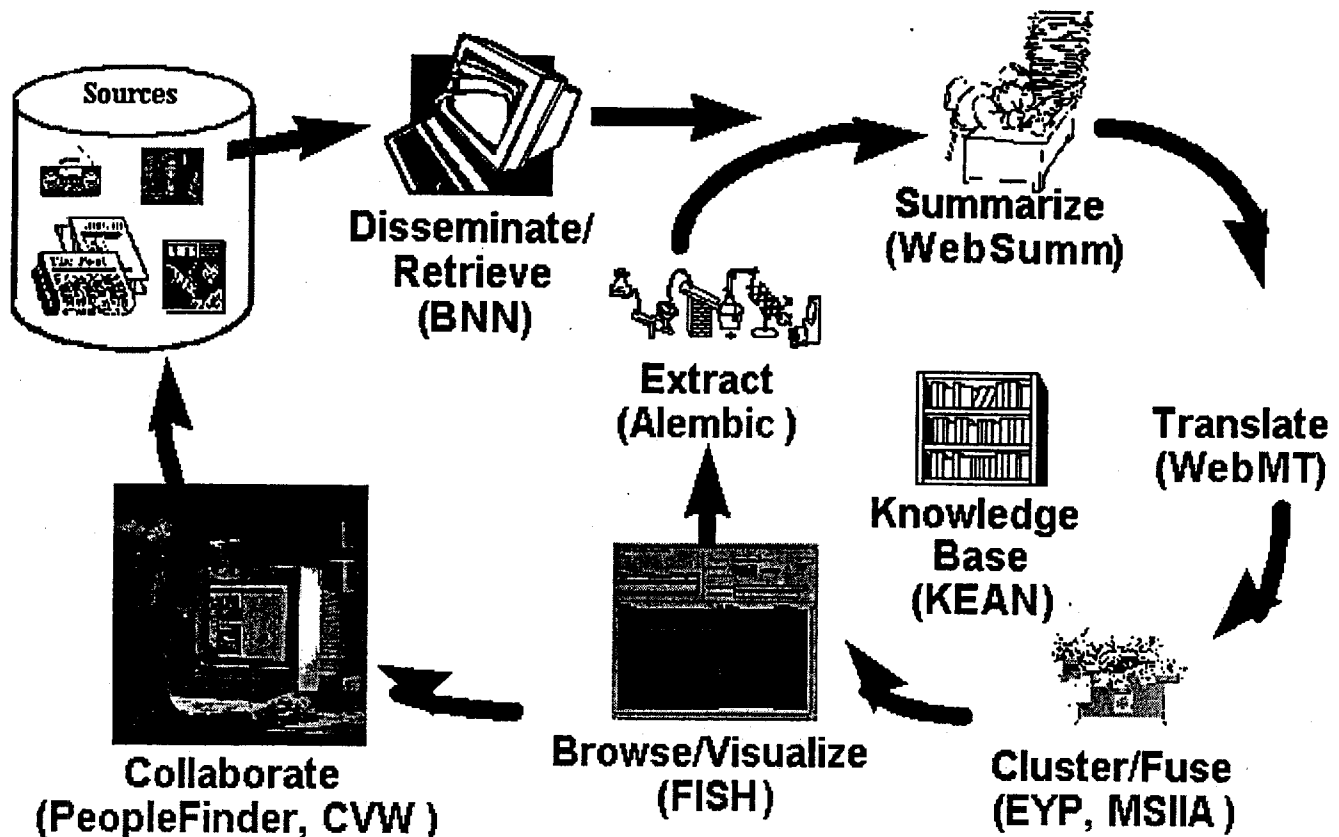
173

Figure 1. Overview of Analytic Process and Tools

As Figure 1 illustrates, a range of tools are applied in this scenario including tools for access to multimedia content (Broadcast News Navigator), information extraction (Alembic), multilingual content (WebMT), multidocument summarization (Websumm), expert discovery (Peoplefinder), multiparty knowledge management (KEAN) and multiparty collaboration (J-CVW). Included in our suite of analytic tools but not illustrated in this specific scenario are tools to cluster information (e.g., Electronic Yellow Pages - EYP) and/or fuse information (e.g., Multisource Intelligence Information and Analysis (MSIIA)), and/or visualize large information collections (e.g., Forager on the Information SuperHighway (FISH) (Smotroff, Hirschman, and Bayer, 1995)).

2. Scenario

Imagine I arrive at the office one morning to discover an urgent request from a Senator for information on Gulf War and suspected proliferation of chemicals from Western industrial corporations. Unfortunately, my top chemical experts are currently away, serving on the US negotiation team for the international chemical weapons treaty. To further complicate matters, the Senator wants all material within 2 hours in order to prepare for a press conference this evening and possible questions that might arise. To make matters even more difficult, all materials must remain unclassified given the public nature of the inquiry. I call an urgent staff meeting. My challenges include creating a timely and informative report without my normal experts using unclassified, heterogeneous, and unstructured sources. After directing several analysts to search our classified holdings, I begin contemplating how I will deal with the mass of multimedia, multidocument, and multilingual unclassified material.

3. Content-based Multimedia Access: Broadcast News Navigator

One of my analysts mention that CNN had a recent report on Gulf War, chemicals and Iraq. We sit down at my web browser and pull up MITRE's Broadcast News Navigator (BNN)™. Every evening at MITRE, BNN digitizes selected broadcast news sources into an MPEG-1 stream and applies image, audio, and (closed caption) text processing algorithms in parallel to eliminate commercials, segment individual stories, extract named entities (i.e., people, organizations, and locations) and select a one-line summary from each story (Maybury, Merlino and Morey, 1997; Mani et al. 1997). Figure 2 shows the most

frequent named entities that have appeared in CNN Prime News over the past month, from which we can detect a mid-month surge in stories mentioning China (reflecting recent US-China negotiations) and an increase in stories about Iraq toward the end of the month (a result of UN chemical/biological inspectors being ousted from Iraq). These are automatically extracted from closed captions using MITRE's Alembic system (Aberdeen et al. 1995), a trainable, corpus-based language processing system that applies the same general error-reduction learning approach used previously for generating part-of-speech rules designed by Brill (1993, 1995) to the problem of learning phrase (i.e., people, organization, location) identification rules. From this display in Figure 2, we can select any of these terms (e.g., Iraq, White House) and retrieve associated stories.
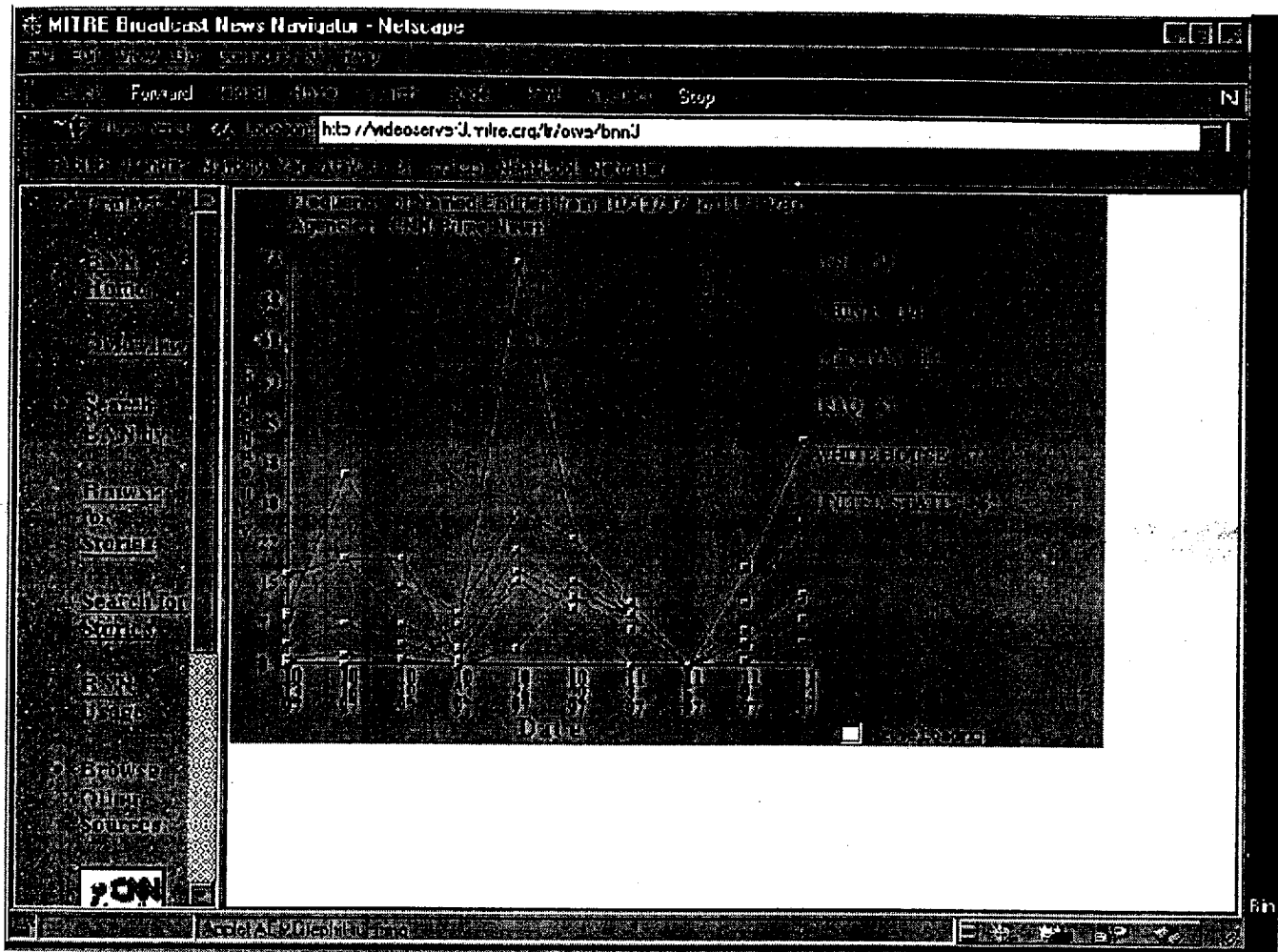
Figure 2. Broadcast News Navigator visualization

Alternatively, we can perform a more refined query. For example, in Figure 3a we use BNN to search for all stories appearing over the past three months in several broadcasts (CNN Early Prime, Headline News, International and Prime News) containing the topic "chemicals", the person "Saddam Hussein, "the location "Iraq," the organization "Pentagon," or the phrase "chemical weapons." In Figure 3a, a keyframe and the top three most frequently occurring named entities are displayed to the user for quick review, enabling them to scan three months worth of the most relevant stories in seconds. I point to the second keyframe in Figure 3a and am hyperlinked to the detailed story view shown in Figure 3b which for each story displays a key frame, pointers to the video and closed caption sources, a classification of the story as a scrollable list of topics (using Oracle's ConText[TM] engine), the named entities extracted from the story, and a one-line text summary of the story. If we point to the key frame, we launch a digital VCR to view the story. This particular story happens to be about Iraq expelling UN chemical weapon inspectors. Subsequent browsing or more refined queries can quickly discover stories specifically addressing Iraq's chemical weapons capabilities. Within a few minutes we discover several chemical compounds are in use by the Iraqi's (e.g., Tabun) as well as mention of a chemical storage site (Kamishiya) destroyed after the war by the US Army. This kind of content based search enables the user quickly review a large collection of broadcast news video to get at who did what to whom, when, where, and how.

175

MITRE Broadcast News Navigator - Netscape

Back Forward   Stop   N

Broadcast Agency(s): CNN Early Prime, CNN Headline News, CNN International, CNN Prime News
From Date: 20-AUG-1997 **To Date:** 18-NOV-1997
Topic(s): chemicals
Person(s): SADDAM HUSSEIN
Organization(s): PENTAGON
Location(s): IRAQ
Text Search: CHEMICAL WEAPONS
Number Of Hits: 139

- Project

BNN
Home

Overview

- Search
BNN by:

Browse
for Stories

Search for
Stories

BNN
Usage

- Browse
Other
Sources:

Story Index for Story(s): 1 to 10

CNN Prime News 14-NOV-1997	CNN Early Prime 12-NOV-1997	CNN Headline News 03-NOV-1997	CNN Prime News 13-NOV-1997	CNN Headline New 03-NOV-1997
IRAQ SADDAM HUSSEIN U.S.	IRAQ SECURITY COUNCIL CNN	IRAQ U.N. BAGHDAD	IRAQ BAGHDAD SADDAM HUSSEIN	IRAQ ISRAEL BAGHDAD
CNN Early Prime 13-NOV-1997	CNN Prime News 14-NOV-1997	CNN Prime News 30-OCT-1997	CNN Prime News 03-NOV-1997	CNN Prime News 12-NOV-1997

176

Summary   **Closed Caption**   **Source**   **Topics**

Key Frame

Video

Related Web Sites

Most Frequent Named Entities

Figure 3a. BNN Story Display Figure 3b. Detailed Story Display

4. WebSumm

Impressed with the ability to quickly identify relevant unclassified video, I remain concerned that I have not identified any detailed content about organizations selling chemicals to Iraq. A query from Alta Vista on "chemical weapons" yields an overwhelming (1,642,608) number of hits, the highest ranked ones of which are concerned principally with the chemical weapons convention (which only frustrates me given my best analysts are currently away supporting this activity). As Figure 4 illustrates, refining the query to "Gulf War" AND "chemical weapons" AND "proliferation" yields little improvement (34,862 hits) as AltaVista is a very fast retrieval engine precisely because it performs no sophisticated statistical or linguistic analysis of documents.

Figure 4. AltaVista Query

I therefore access MITRE's WebSumm (Mani and Bloedorn 1997) system via my web browser and submit the same search (ìGulf Warî and ìchemical weaponsî) which yields the results shown in Figure 5. Like AltaVista, a user can choose the number of hits to retrieve and alter the way in which they are displayed (e.g., number per page, sort order, and sort by date, size, title, or rank). Indeed, WebSumm initially uses AltaVista to retrieve articles from the web which it then subsequently analyzes using statistical and linguistic processing, for example selecting key sentences most relevant to the user's query to describe each article. In the left hand pane of Figure 5, the system has computed the most frequently occurring terms across all documents (called "shared content terms"), listing this metadata so the user can obtain a quick assessment of the information space. The user can then promote or demote each term to rapidly and iteratively re-rank the document list. For example, if we were more interested in documents involving biological as opposed to chemical weapons, we could neutralize terms like ìchemicalî and ìchemical weaponsî and increase words like "biological" and ìnerve gasî, which would reorder the document list.

The document pane on the bottom right hand side of Figure 5 lists the rank-ordered documents, including the location they appeared in the original AltaVista list (e.g., the first document appeared 48th in AltaVista's returned list). WebSumm re-ranks the retrieved documents using the user's query and lexical knowledge from an on-line thesaurus (WordNet) so that more relevant documents are placed higher in the document list. Each document is described by its headline and key lines extracted that are most relevant to the query. A user can choose to summarize a document in several ways: extracting a specified number of key sentences, extracting fragments, or returning matching keywords in context. In our example, if the user chooses the third document from the list (doc #5), a key sentence summary quickly reveals "mustard gas, CS gas, tabun and sarinî as important chemical weapons. Another document in the first few documents (#15) specifically addresses a clamp down on shipments to Iraq of precursor chemicals. Indeed, one of the extracted summary lines (line #21 of document #15) indicates that the "State and Commerce departments are now restricting exports of five chemicals: potassium fluoride, dimethyl methylphosphonate, methylphosphonyl difluoride, phosphorous oxychloride and thioglycol" because these can be used in the production of chemical weapons. The article further notes

Figure 5. WebSumm Results on "Gulf War" and "chemical weapons" query

the common use of pesticide production as a legitimate cover for illegal purchases of these chemicals. We soon discover from document #64 (related to #75) that 84 German firms and 18 US/British firms were implicated in illegal chemical sales. In addition to aiding in the rapid discovery and analysis of individual documents, Websumm enables the user to pinpoint similarities and differences between documents, generate composite summaries from multiple documents, and align text segments to enable comparative analysis of documents. Following each document in the document list in Figure 5 is a list of documents most similar to it. This enables the user to pick similar documents to compare to one another as shown in Figure 6. In Figure 6 the user has asked the system to compare documents #48 and #14 and quickly discovers two suspected West German (chemical) companies: Karl Korb and Sigma Chemie.

179

The Times said that a West German scientific supply company, Kar F...b, arranged to sell the pesticide production equipment that will be used to manufacture the weapons.

The company denies this.

The State Department confirms that it is talking to West Germany about the possibility that such equipment 'may have been shipped' to Iraq, and a spokesman says the Germans are investigating on an "urgent basis."

Officials would not comment for the record on the nature of the Iraqi weapons program on grounds that doing so would compromise intelligence sources.

However, one policy-maker and expert in the chemistry of gas weapons confirmed the thrust of the Times story in a telephone interview with Science.

He said it may be an exaggeration to say there are five deeply buried gas factories.

But 'there may be five buildings."

Of the German firm's denial, he said, "Don't trust it."

The U N paper contains the most substantial evidence in the case

Hassan al-Majid, who emerged last month at Jordan.

The existence of Yellow-Rain-type weapons in the Iraqi arsenal is among a number of startling disclosures, including the existence of enough anthrax in the Iraqi stockpile to kill every man, woman and child in the world, success in making weapons from fungal toxins and the existence of large-scale poison factories completely undetected by United Nations inspectors.

Bush Threats Feared

Yellow Rain bombs, along with anthrax and botulinim toxin bombs and warheads, were, according to U N monitors, deployed at the end of December of 1990, in the run-up to the American-lead ground attack that drove the Iraqis from Kuwait.

The weapons were removed from the battlefield by the Iraqis only after Baghdad received warnings from President Bush and Secretary of State Baker that they interpreted as a threat to use nuclear weapons in retaliation.

Iraq now claims it destroyed the weapons after the Gulf War, but U N monitors have found no

Figure 6. WebSumm comparison of similar documents

5. People Finder

Our analysis up to this point has identified types of chemical weapons, export bans on chemical precursors, and a few errant Western companies. To accelerate our efforts, because our local experts are away, we would like to identify corporate experts in chemical weapons and pull them into our task via a desktop collaboration. We fire up MITRE's PeopleFinder (Mattox, Smith and Seligman 1998) and search on îchemicalî as shown in Figure 7. PeopleFinder ranks employees by the number of mentions of a term or phrase and its statistical association with the employee name either in corporate communications (e.g. newsletters) or based what they have published in their resume or document folder (a shared, indexed information space). Integrated with our corporate employee database, employees are ranked by frequency of mentions, pointing to sources in which they appear. Decreasing project size, distribution of staff, and cost/time pressure are driving a need to leverage enterprise expertise and quickly discover who knows what to find and form teams, a gap which People Finder nicely fills.

180
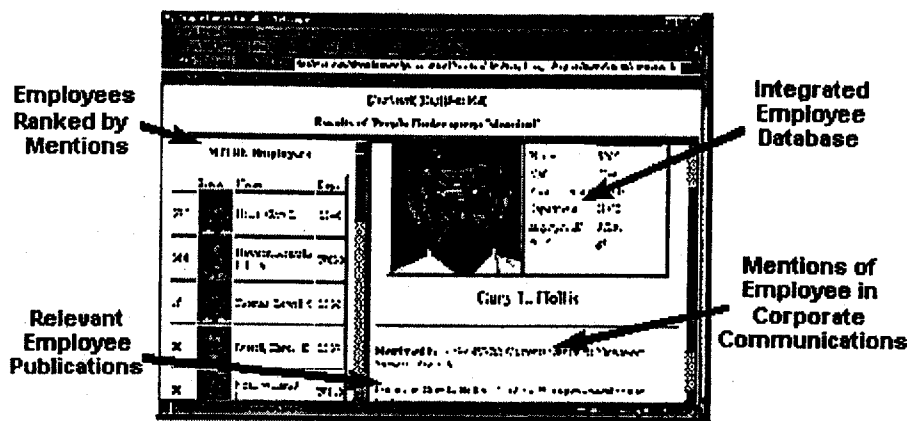
Figure 7. People Finder

6. Collaborative Environments

While examining documents associated with the above experts we focus in on the third expert, Lowell Thomas, and decide we would like to interact with him. We log in via our web browser to MITRE's Java-based Collaborative Virtual Workspace (J-CVW) (Spellman et al. 1997). As shown in Figure 8, from the user perspective CVW includes an integrated suite of facilities that enable synchronous and asynchronous collaboration, including text chat, audio and video conferencing, shared whiteboard, shared and private data spaces, and persistence of sessions via recording user interactions as they occur within the context of shared virtual rooms. CVW distinctively provides location independence and transparency as well as a room metaphor to support context management. When we log on, we ask to "join" Lowell and are transported to his virtual location. When we ask him if he has any knowledge about suspected Germany companies that might be selling chemicals, Lowell suggests searching the Internet for "chemie" and "gmbh".
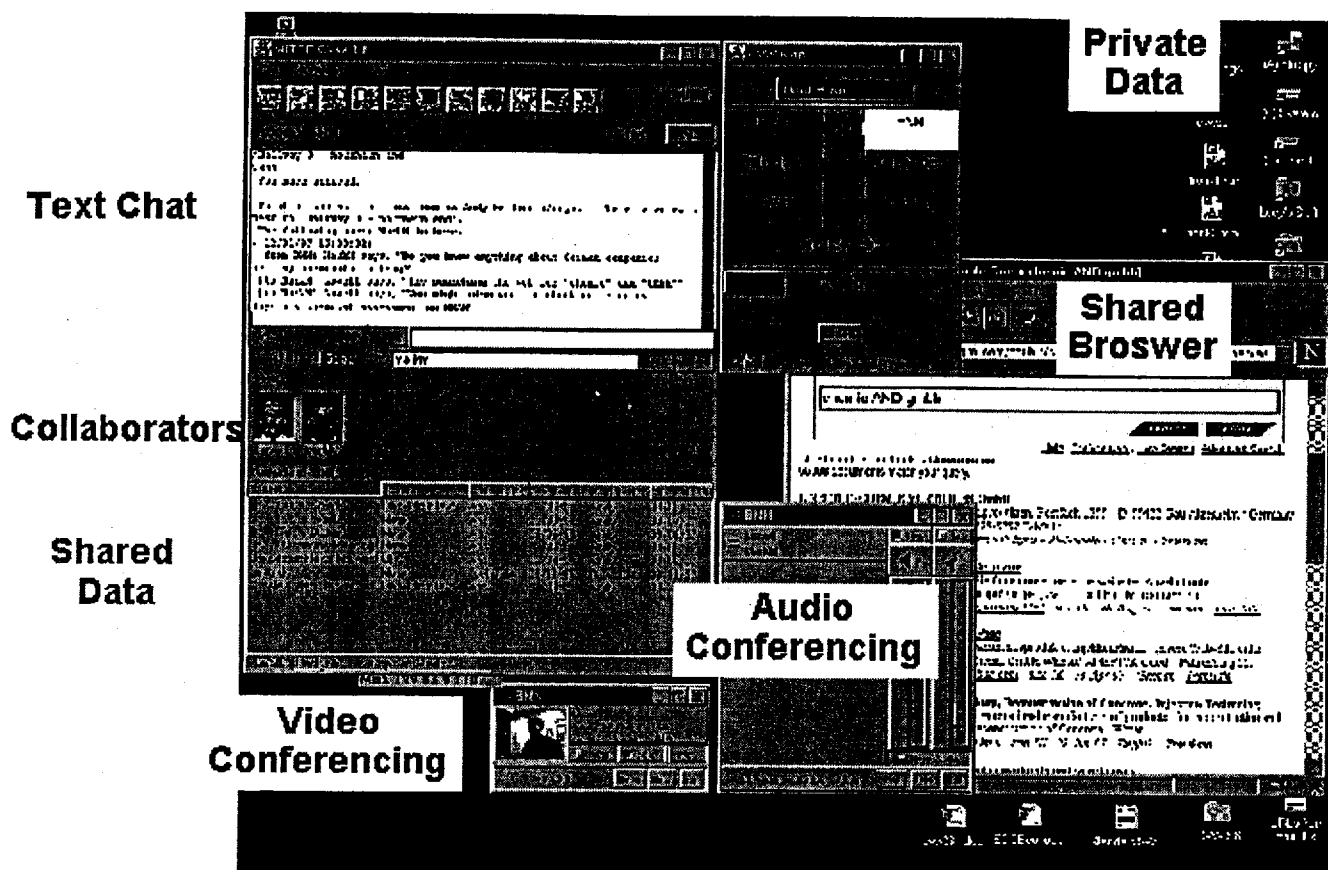
Figure 8. Java Collaborative Virtual Workspace (J-CVW)

7. WebMT

Figure 9 illustrates such a search using Alta Vista. We obtain several interesting sites. Unfortunately, not all are in English. We select one in particular as shown in Figure 10a. While this site is entirely in German we simply access WebMT, a web-based version of Systran's machine translation software developed for the National Air Intelligence Center, paste in the URL, and hit the Translate button. This results in Figure 10b, in which we have preserved the original HTML markup, but have used Systran to provide "gist" quality translations of the content. While not perfect, these translations are sufficient to enable an analyst to make relevancy assessments and perhaps to discover mentions of the export-restricted chemicals identified above.
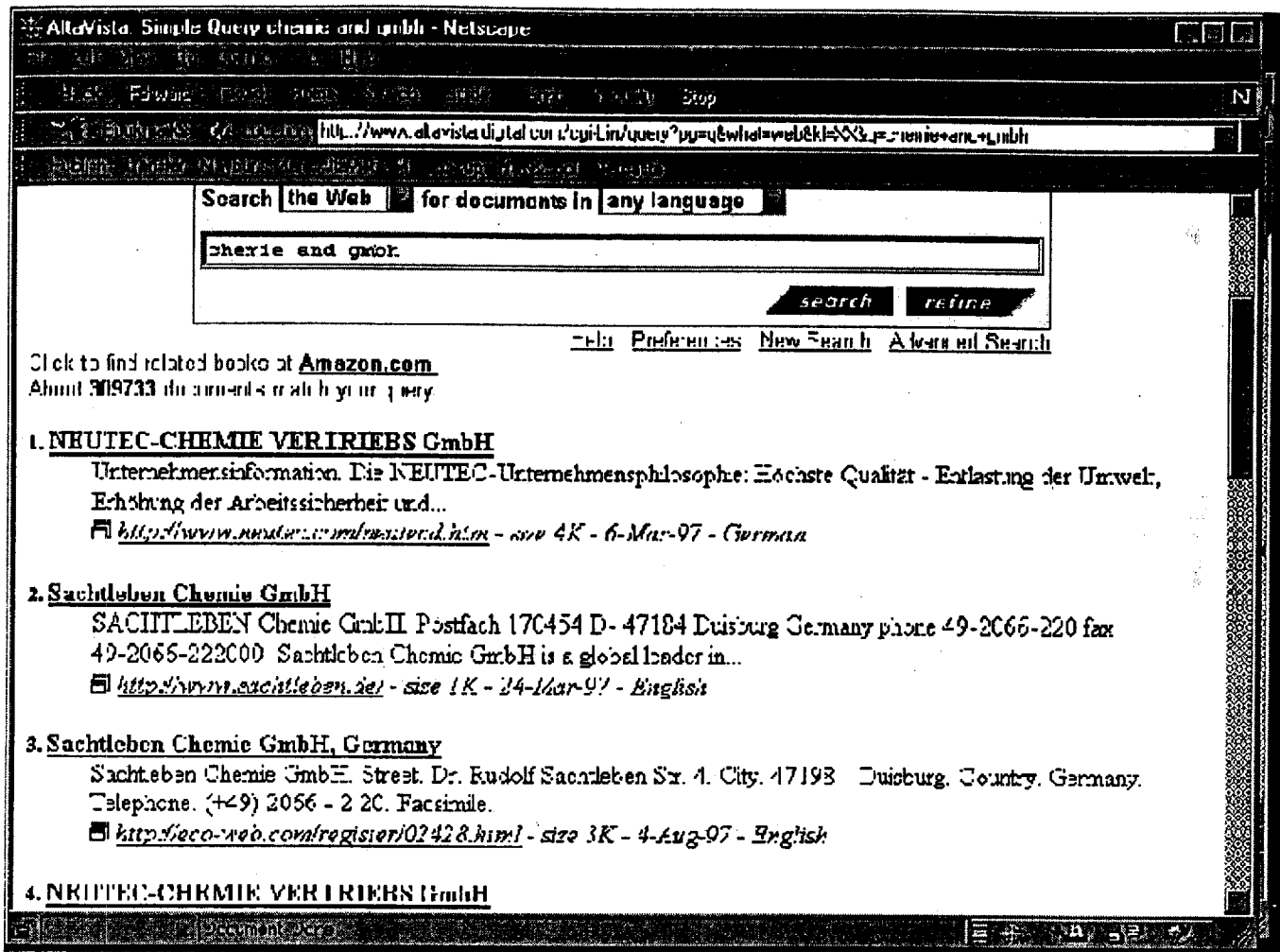
Figure 9. Foreign language pages on the Internet
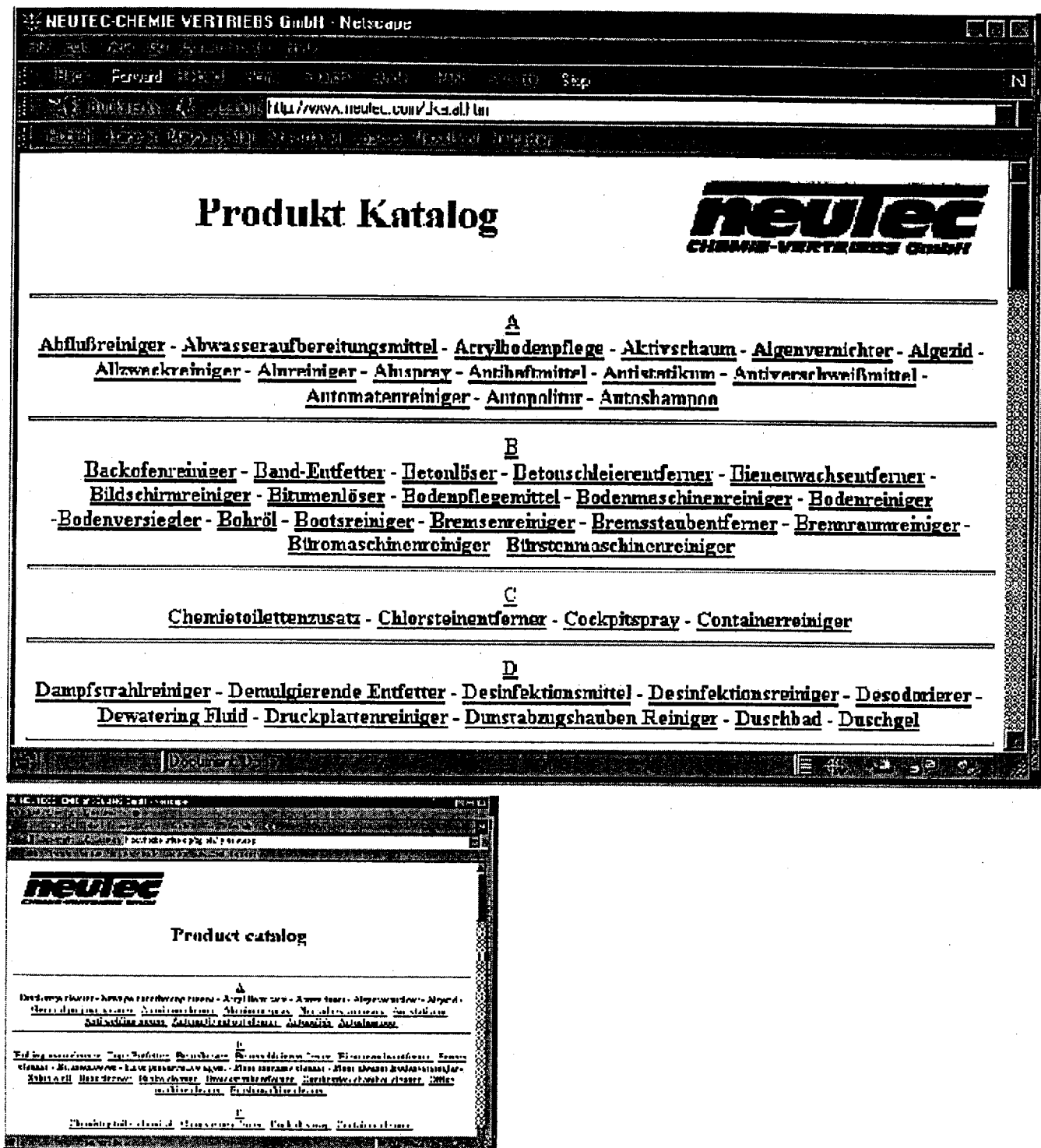
Figure 10. A. Original Foreign and B. Translated Language Internet Page

8. KEAN - Knowledge Evaluation and Annotation

Our expert, Lowell, recontacts us in CVW and suggests using MITRE's knowledge management tool, KEAN, to investigate what one of our traveling chemical experts, Stanley Boykin, has said about chemical assessments. We bring up KEAN as shown in Figure 11a and search on "boykin" and "chemical" and obtain the results as shown in Figure 11b. KEAN acts like a window onto the web, for each source listing individuals assessments (rankings from 1-10, classifications of content into

184

pre-specified categories), as well as listing the most related sites. We quickly scan these sites and discover University of Sheffieldís ChemDex (http://www.shef.ac.uk/~chem/chemdex/), a web site that catalogues academic, industrial, and government organizations involved in research or manufacture of chemicals. This immediately leads us to a detailed list of chemical manufacturers. We look up the German (Figure 12) and American ones and continue our analysis (e.g., which manufacture restricted exports, which have offices in Iraq) to come up with a more refined suspect list.

185

Figure 11. A. KEAN Search and B. Source, Assessment, Related Pages

Figure 12. German companies in Sheffieldís Chemical Index (ChemDex)
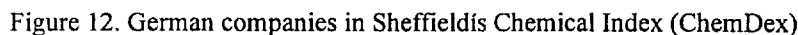
9. Summary

While we still have more work to do, in a very short time, we have browsed months worth of broadcast news, visited thousands of sites, and summarized, translated and extracted key information to discover export controlled precursor chemicals, past illegal German sales to Iraq, and potential suspect companies for further analysis. Using desktop collaboration, we identified German chemical manufacturing companies, visited their foreign language home pages, and translated content to evaluate the relevancy of advertised products. In further synchronous collaboration with chemical experts and through use of knowledge management tools, we were directed to a key list of chemical manufacturers. All of this analysis was done from our desktop via web-based interfaces.

Introduced in Figure 1, the demonstrated tools extend commercial capabilities in a variety of important ways. For example, BNN extends commercial relational database, web, and video server technology to enable broadcast news story segmentation, named entity extraction, summarization and visualization. Alembic has enabled us to extract information from unstructured sources. WebSumm extends AltaVista's web search by extracting and displaying shared content terms, supporting post-retrieval reordering, and summarizing documents. WebMT enables users Web-based access to Systran's gist quality machine translation for Web-page translation. KEAN has managed the group evaluation and annotation of content. While not described, separate efforts at MITRE are focused on metadata management (EYP), multisource information analysis and fusion (MSIIA), and large scale document visualization (FISH). We have further shown the ability to discover expertise (PeopleFinder). Finally, using this knowledge we launch desktop collaboration (CVW) to provide integrated access to audio and video conferencing (Lawrence Berkley Labs VIC and VAT) as well as providing persistence and shared context via mechanisms such as virtual rooms. Taken together, these capabilities illustrate integrated access to tools to exploit multimedia, multidocument, multilingual, and multiparty information and knowledge.

10. Future Vision

In conclusion, Figure 13 outlines areas we believe are critical to the future analytic environment. As illustrated, we believe

*18*7

collaboration environments, knowledge management, performance support, intelligent agents, multimedia/multilingual digital libraries (Maybury 1997), and intelligent human computer interaction (Maybury 1993) are all key to the success future information analysis tasks. Important additional requirements include mechanisms to ensure secure collaboration, including the need to extend interactions across security levels and to external experts. This paper reports on several efforts to provide tools that enhance an analyst's ability to rapidly create knowledge from heterogeneous information sources and distributed expertise.
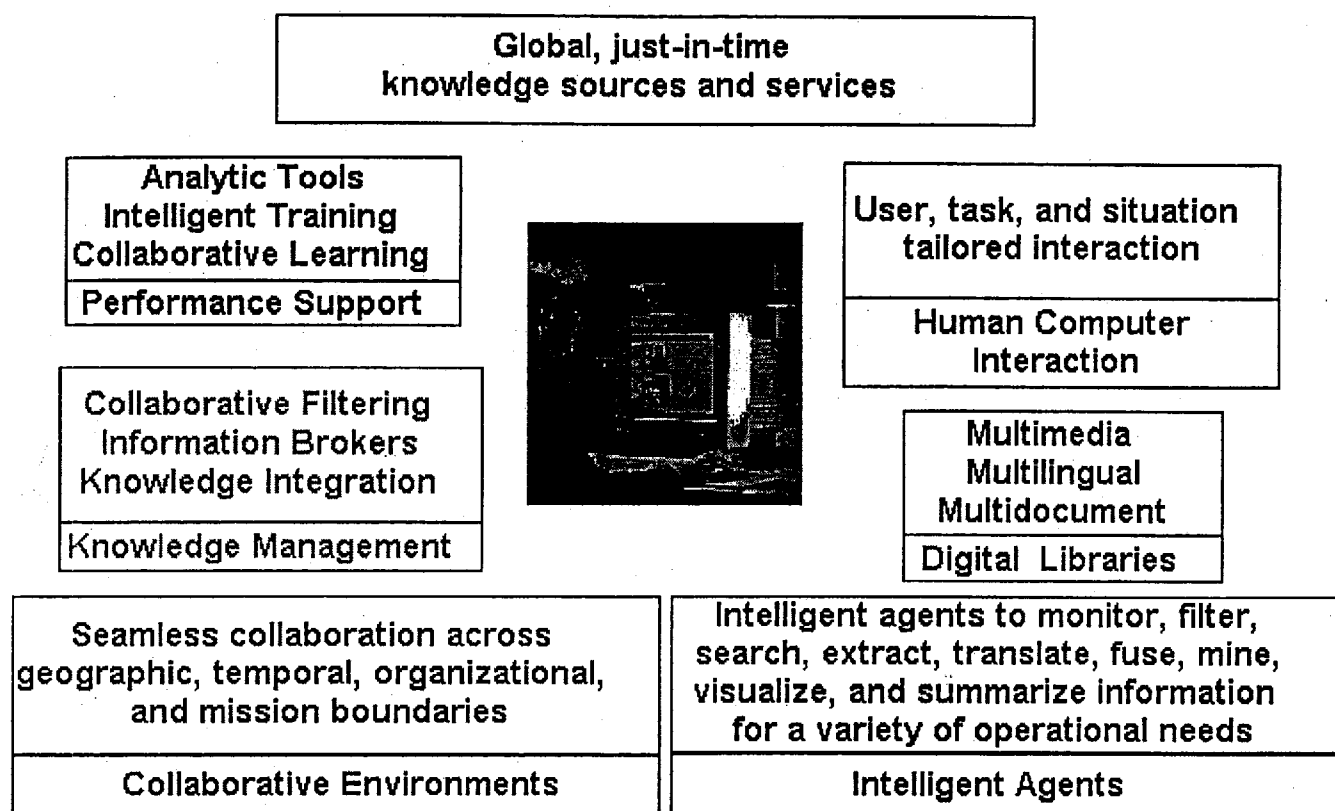

Figure 13. Vision of Future Analytic Environment

A number of extensions to the above tools are underway such as scaling BNN to automatically process multiple channels or determining the optimal summary to generate for a given source, user, and purpose. Another area of active research is taking the output from information extraction (i.e., Alembic) and mining these for trends and unusual patterns. For example, Clifton et al. (1997) and Tsur et al. (forthcoming) use data mining tools to investigate three-way associations between two people and a location in 6,612 news stories, and automatically correlated two individuals Mobutu Sese Seko and Laurent Kabila (essentially the former and current presidents) with Kinshasa (current capital of Congo). Other important areas for continued research include synergistic interaction between this class of analytic tools and between the tools and the user. For example, while the web provides a common mechanism for tool access, tighter integration among the tool's inputs and outputs is desired; e.g., taking search results and importing these to translation, then extraction and/or summarization routines. Also, the current interface is primarily passive, whereas with the integration of more intelligent tools for query and data analysis (e.g., mining), we expect the need for more active, agent-based systems that engage the user in a mixed-initiative dialogue to solve their problems.

11. Acknowledgments

Andy Merlino is responsible for the Broadcast News Navigator. Inderjeet Mani, Eric Bloedhorn and David House designed and developed WebSumm. Dave Mattox developed PeopleFinder. Peter Spellman and colleagues developed CVW. Ira Smotroff, Ann Tallant and colleagues created FISH. Marc Vilain and John Aberdeen provided part of speech and proper name taggers, and David Day for training these on the broadcast news domain. Rob Zakon developed WebMT for the National Air Intelligence Center using Systran's machine translation engine. KEAN was developed by Daryl Morey. CNN is a registered trademark of Cable News Network.

12. References

Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P., and Vilain, M., 1995. iDescription of the Alembic System Used for MUC-6," Proceedings of the Sixth Message Understanding Conference. Advanced Research Projects Agency Information Technology Office, Columbia, MD, 6-8 November 1995.

188

Brill, E. 1993. A Corpus-based Approach to Language Learning. Doctoral dissertation, University of Pennsylvania.

Brill, E. 1995. Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics* 21(4): 543-565, December.

Clifton, C., Rosenthal, A. and Ullman, J. D. 1997. Knowledge Discovery in Text. In the First Federal Data Mining Symposium, AFCEA, December 16-17, 1997, Washington, D.C.

Mitchell, R., Day, D., and Hirschman, L. 1995. Fishing for Information on the Internet," submitted to *IEEE Symposium on Information Visualization (Info Vis '95)*, Atlanta, October 1995.

Mani, I. and Bloedorn, E., 1997. Summarizing Similarities and Differences among Related Documents. In Proceedings of the Fourteenth National Conference on Artificial Intelligence, p. 622-628. AAAI Press, Menlo Park, CA

Mani, I., House, D., Maybury, M. and Green, M.,1997. Towards Content-Based Browsing of Broadcast News Video. In Maybury (ed.) *Intelligent Multimedia Information Retrieval*. AAAI/MIT Press.

Mattox, D., Smith, K., and Seligman, L. 1998. Software Agents for Data Management. In Thuraisingham, B. *Handbook of Data Management*, CRC Press: New York. 703-722.

Maybury, M. T. (ed.), 1993. *Intelligent Multimedia Interfaces*. Menlo Park: AAAI/MIT Press. (http://www.aaai.org/Publications/Press/Catalog/maybury.html).

Maybury, M. T. (ed.), 1997. *Intelligent Multimedia Information Retrieval*. Menlo Park: AAAI/MIT Press. (http://www.aaai.org:80/Press/Books/Maybury-2/).

Maybury, M., Merlino, A., and Morey, D., 1997. "Broadcast News Navigation using Story Segments", ACM International Multimedia Conference, Seattle, WA, November 8-14, 381-391.

Morey, D. 1997. Enterprise Knowledge Management. In Thuraisingham, B. (ed.) *Data Management Systems*, CRC Press: New York, 723-729.

Spellman, P. J., Mosier, J. N., Deus, L. M., Carlson, J. A., 1997. Collaborative Virtual Workspace. Proceedings of International ACM SIGGROUP Conference of Supporting Group Work, 16-19 November, Phoenix, Arizona,197-203. ACM Press: NY.

Smotroff, I., Hirschman, L., and Bayer, S. 1995. "Integrating Natural Language with Large Dataspace Visualization," to appear in Adam, N. and Bhargava, B. (eds), Advances in Digital Libraries, Lecture Notes in Computer Science, Springer Verlag, 209-224.

Tsur, S.; Ullman, J. D., Abiteboul, S., Clifton, C. Motwani, R. Nestorov, S. and Rosenthal, A. "Query Flocks: A Generalization of Association Rule Mining" Submitted to the 1998 ACM SIGMOD International Conference on Management of Data, 1-4 June, 1998, Seattle, WA. http://www-db.stanford.edu/pub/papers/flocks.ps.

189

PROCEEDINGS 1998 7th International Conference & Exhibit OPEN SOURCE SOLUTIONS: Global Intelligence Forum - Link Page

Return to Electronic Index Page