PERSONALIZED MULTIMEDIA INFORMATION ACCESS

[MARC LIGHT AND MARK T. MAYBURY]

Ask questions, get personalized answers.

Increasing sources and amounts of information challenge users around the globe. Newspapers, television news broadcasts, Web sites, and other forms of communication provide society with the vast majority of real-time information. Unfortunately, cost and time pressures demand that producers, editors, and writers select and organize content for stereotypical audiences. Here, we illustrate how content understanding, user modeling, and tailored presentation generation offer the potential of personalized interaction on demand. We describe two systems that represent components necessary for an adaptive question answering system: a nonadaptive question answering system and another for tailored news video retrieval.

We are investigating systems that will allow users to perform searches by asking questions and getting personalized answers rather than by typing in keywords and typically getting an overwhelming number of often irrelevant Web pages. This requires advancing beyond the conventional information retrieval strategy of document/Web page access toward the automated extraction and summarization of possibly multilingual and multimedia information from structured, unstructured, and semistructured sources followed by the generation of a personalized presentation. Taken together, the two systems we describe represent steps toward our vision of personalized question answering.

Automated question answering is the focus of the Text Retrieval Evaluation Conference (TREC) Question Answering track and the ARDA Advanced Question Answering for Intelligence (AQUAINT) [1] program. Question answering differs from more traditional forms of information retrieval in that it aims to provide answers to queries as opposed to documents. For example, Breck et al.'s Qanda (Question AND Answering) [2] system aims to find explicitly stated answers in large document collections. For example, a question such as "Who was the architect of the Hancock building in Boston?" posed against a collection of five years of the *Los Angeles Times* might result in retrieving the statement "I.M. Pei was a student at M.I.T ... He designed the John

Hancock building in Boston." The majority of research in question answering uses unstructured textual documents as the source of answers; however, work on more structured sources is also being performed. Qanda's research aims are to:

- Understand the question well enough to "ask" the knowledge sources by extracting the characteristics of the answer and routing the question to the best knowledge source.
- Understand the knowledge sources well enough to find the answer whether it is in a relational database, semistructured data, or human language text without structure.
- Discover how this capability can help fulfill a user's information need, which could include ad hoc questions against a static database, standing questions against a stream of data, a series of related questions, or templated questions.

The processing flow in Qanda follows a number of key stages. The first stage is question analysis, which determines the type of object that answers the question (for example, "when" needs a time, "where" a place). In a parallel second stage, relevant documents are retrieved using keywords from the question. Stage three is document processing, in which retrieved documents are searched for entities of the desired type using information extraction. In stage four, the candidate answers are ranked. Finally, in stage five, the best forms of the candidate answers are chosen (for example, "I.M. Pei" instead of just "Pei").

Consider an example of Qanda's performance for the question "Where did Dylan Thomas die?" The system responds:

✖ 1. Swansea: In *Dylan: The Nine Lives of Dylan Thomas*, Fryer makes a virtue of not coming from **Swansea**.

✖ 2. Italy: Dylan Thomas's widow Caitlin, who died last week in **Italy** aged 81.

✓ 3. New York: Dylan Thomas died in **New York** 40 years ago next Tuesday.

In this example the first two answers are incorrect but the third is correct. Qanda provides a rank ordered list of answers, highlighting the extracted location from a source document and including a text fragment from the source in which the location occurred.

Question Answering Performance

The TREC-8 Question Answering track evaluation [10] involved 25 participating systems. Each system was given 198 questions to which it had to find answers from a collection of approximately one-half million documents. Each system run consisted of up to five ranked answers per question. Two separate runs were allowed, one consisting of responses 50 bytes or shorter and one consisting of responses 250 bytes or shorter. Systems were compared using mean reciprocal rank. "Rank" here refers to the rank of the first correct answer returned by a system. The reciprocal is one divided by this rank. For example, if the first two answers provided by a system were wrong but the third was correct, the reciprocal rank would be 1/3. The mean reciprocal rank is the average over the 198 questions. As illustrated in Figure 1, the top performing system had a 0.66 mean reciprocal rank. This system had the correct answer among its top 5 responses for 73% of the questions. Figure 1 also shows those systems that used natural language processing as darker colored bars and those that used word-based information retrieval as lighter colored bars. The Qanda system, shown in the middle of the graph in the darkest color, uses language processing.

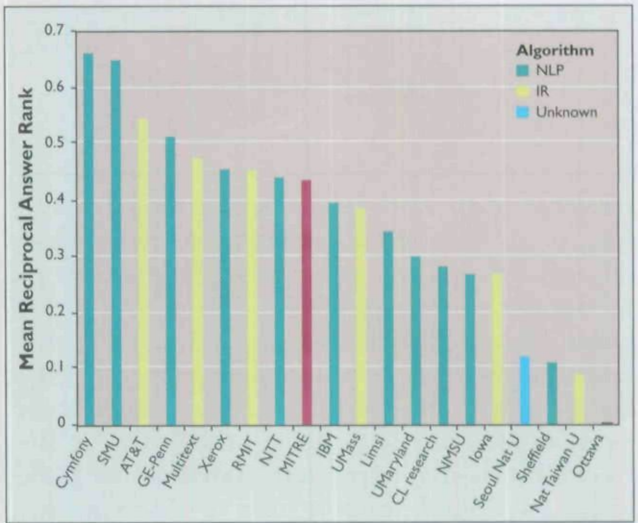

Figure 1. Question answering system performance.

Toward Multimodal Question Answering

Extensions to question answering systems like Qanda focus on multimedia and multilingual sources. Content-based access to multimedia (text, audio,

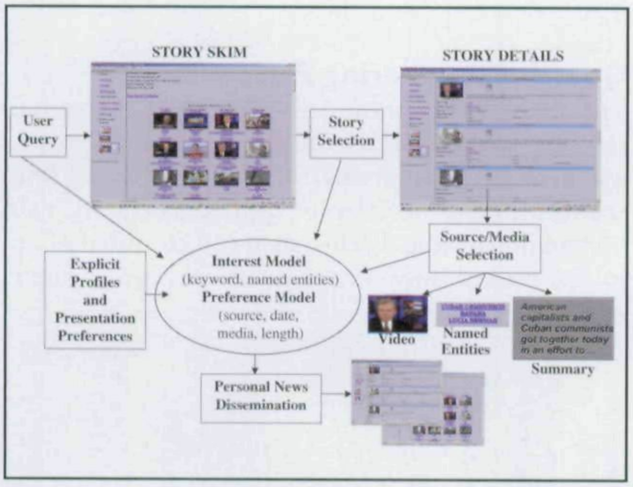
STORY SKIM
STORY DETAILS
User Query
Story Selection
Source/Media Selection
Explicit Profiles and Presentation Preferences
Interest Model
(keyword, named entities)
Preference Model
(source, date, media, length)
Video
Named Entities
American capitalists and Cuban communists got together today in an effort to...
Summary
Personal News Dissemination

Figure 2. Story skim, story details, and personalization of Cuba stories.

imagery, video) and multilingual content promises on-demand access to sources such as radio and broadcast news tailored to a range of computing platforms (for example, kiosks, mobile phones, PDAs). Today, people are daily offered vast quantities of news in the form of multiple media (text, audio, video). For the past several years, a community of scientists has been developing news on demand algorithms and technologies to provide more convenient access to broadcast news [6]. Systems have been developed that automatically index, cluster/organize, and extract information from news. Synergistic processing of speech, language, and image/gesture promise both enhanced interaction at the interface and enhanced understanding of artifacts such as Web, radio, and television sources. Coupled with user and discourse modeling, new services become possible such as individually tailored instruction, games, and news ("personalcasts").

Broadcast News Navigator

To illustrate personalcasting, we briefly describe our research to create the Broadcast News Navigator (BNN) system, which exploits video, audio, and closed caption text information sources to automatically segment, extract, and summarize news programs [8]. BNN can also automatically learn segmentation models from an annotated video corpus [2]. The Web-based BNN system gives a user the ability to browse, query (using free text or named entities), and view stories or their multimedia summaries. For each story, the user is given the ability to view its closed caption text, named entities (for example, people, places, organizations, time, or money), a generated multimedia summary, or the full original video of a story. For example, Figure 2 shows the results of BNN's response to a user query requesting all reports regarding Cuba between May 17 and June 16, 2001. In the presentation in Figure 2 called "story skim," for each story matching the query, the system presents a key frame, the three most frequent named entities within the story, and the source and date of the story.

This, in essence, provides the user with a "Cuba" channel of information from multiple broadcast news sources, personalizing the channel to his or her information interests. Moreover, the user can create arbitrarily complex queries combining keywords, named entities (people, organizations, locations), source (CNN, MSNBC, ABC), or time intervals (specific days, weeks, or years). These queries result in selected video stories specific to the user's interest.

The user can then select any of the key frames in the "story skim" to get access to details of the story, such as shown in Figure 2 in the view labeled "story details." In this presentation, the user has access to all of the people, organizations, and locations mentioned in the story. At the same time the user has access to an automatically extracted one-line summary of the news (the sentence with the most frequent named entities), a key frame extracted from the story segment, and a pointer to the full closed-captioned text and video source for review.

The system provides navigation support, so that the user can select named entities and find stories that include them. Further, by employing a clustering algorithm, the system enables the user to select stories similar to the current story. When we give the user hypertext access to both "story skim" and "story details," we will call this the "skim and details" method.

Empirical Studies

In order to better understand the value of the various kinds of presentations shown in Figure 2, Merlino and Maybury [9] had 20 users perform relevance judgment and comprehension tasks using various mixes of presentation media. In particular, the user's task was either to decide if stories belonged to a given topic (identification task) or to extract information from the stories such as the people, organizations, locations, and topics mentioned in them (comprehension task). Figure 3(a) plots the average recall of the users in performing the identification task (relevancy judgments) against their precision in doing so. Precision is the percentage of documents selected that are actually relevant whereas recall measures the percentage of documents that were actually selected from all those that

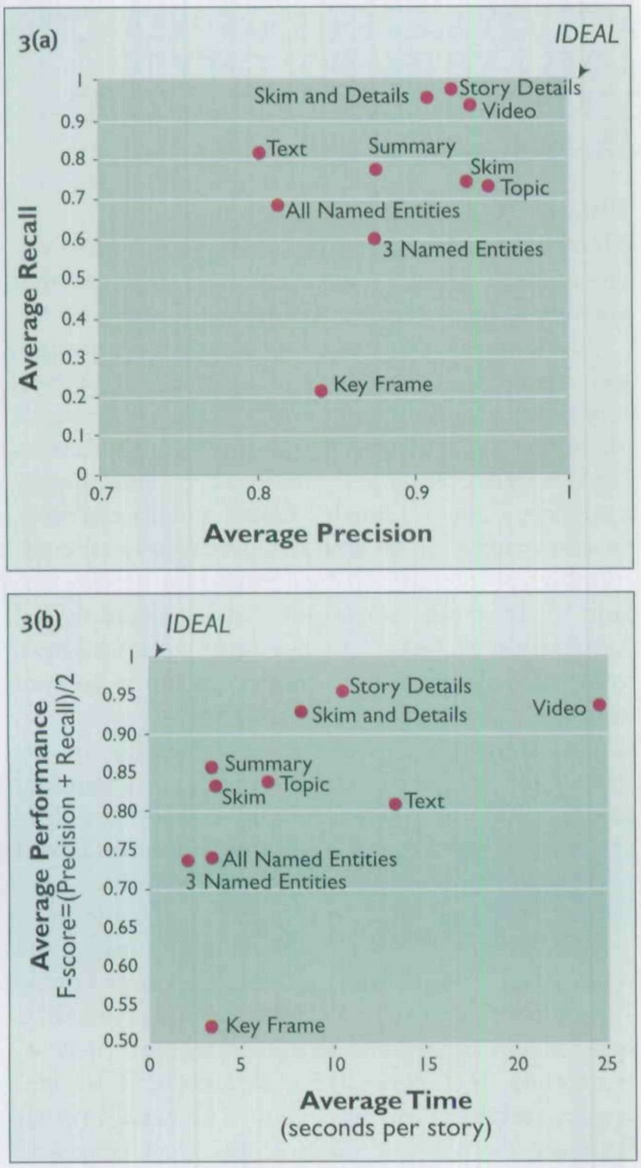

Figure 3(a) Recall versus precision performance for different multimedia displays. Figure 3(b) Performance versus time for different multimedia displays.

should have been selected from the corpus.

Each point on the graph in Figure 3a plots the average performance of all users for a giveng presentation media (for example, key frames, named entities, one-line summaries) or mixes thereof. As is shown, using "story details," "skim and details" and the original source "video" users on average exhibit precision and recall performance above 90%. In contrast, key frames exhibit poor recall, a consequence of lack of information in the key frames themselves and a relatively naive key frame selection algorithm.

Conversely, in Figure 3(b) each point on the graph plots the average performance of all users (F-score—an equally weighted average of precision and recall) versus the average time taken to process each story for different presentation media. What is evident from the graphs is that BNN's "story details" and "skim and details" ("story skim" plus "story details") presentations result in user retrieval performance above 90%. Only "video" has a similar performance level, but that source takes more than twice as long per story. If users can tolerate higher errors (80–85% F scores), then "summary," "topic" and "skim" presentations can further decrease the time per story by as much as a factor of two or three. Finally, we also compared average user performance when the answer data sets were small (less than 10) and large (greater than 10). We found with larger data sets, users exhibit higher precision using the skimming technique as they could more quickly scan larger collections of stories.

In addition to enhanced task performance, users reported higher satisfaction for these two access methods. Using a scale from 1=dislike to 10=like, users indicated a preference for these two methods when identifying relevant stories (a 7.8 average rating for these two methods compared to a 5.2 average for other methods). They also preferred "story details" and "skim and details" methods for comprehension tasks in which the user had to extract named entities from sources (an 8.2 average rating versus a 4.0 average for other methods).

In summary, some of the key findings from the user study included:

- There was no difference in average human retrieval accuracy between most mixed media presentations and original video source.
- Poor source quality (for example, average word error rates of 12.5% in closed caption texts) hindered access performance.
- Presenting less information to the user (for example, skims or summaries versus full text or video) enabled more rapid content discovery.
- Story skims are better for larger data sets.
- Mixed media presentations were the most effective presentation for story retrieval and comprehension.

User Modeling and Tailoring

A typical search session in BNN follows a three-step process. As exemplified in Figure 2, the user first poses a query and receives a story skim. The user then selects a story and is provided the details. From this story detail, the user can simply review the summary and all named entities or explicitly choose a

BECAUSE THE ORIGINAL BROADCAST NEWS SOURCE IS SEGMENTED INTO ITS COMPONENT PARTS, KEY ELEMENTS ARE EXTRACTED AND/OR SUMMARIZED. THIS ENABLES A SYSTEM NOT ONLY TO SELECT STORIES BASED ON USERS' CONTENT INTEREST, BUT ALSO TO ASSEMBLE THEM IN THE MANNER A USER PREFERS.

media element to display, such as the full video source or the text transcript. Each of these user actions affords an opportunity for modeling user interest.

The user interest profile can be created from explicit and/or implicit user input and then used to tailor presentations to the user's interests and preferences. As shown in Figure 2, in BNN the user can explicitly define a user profile by defining simple keywords, or semantic entities such as individuals, locations, or organizations indicating their interest. They can also specify preferred broadcast sources to search (such as CNN or ABC News). This profile has been designed to be extended to indicate media type preferences for presentation (key frame only, full video, text summary), possibly driven by the user's current platform (mobile phone, handheld, desktop), user location, cognitive load, and so on. The user's interest profile is run periodically and sent to the requester as an alert or as story skims or details like those shown in Figure 2.

In addition to this explicit collection of an interest model, we have designed (but not yet implemented) an implicit method to build an interest model by watching the user session to track the user's query, selection of particular stories, and choice of media.

Because the original broadcast news source is segmented into its component parts, key elements are extracted and/or summarized. This enables a system not only to select stories based on users' content interest, but also to assemble them in the manner a user prefers. For example, the user can be presented with only key frames, with summary sentences, with people or place names, or with the entire source. Furthermore, interests and preferences can be related in interesting ways. For example, we can discover which user preferences regarding source, date, length, or media elements correspond to specific keywords, people, organizations, or locations, and vice versa.

A natural extension of this work would be to add a feedback and collaborative filtering mechanism so that not only would the individual user's interest and preference model modify with each search, but also the user could benefit from searches performed by others in a community.

Future Research

Many outstanding research problems must be solved to realize automatically created, user-tailored answers to questions. For example:

Instrumentation of user applications to automatically log and infer models of user interest. With users increasingly learning, working, and playing in digital environments, instrumentation of user interactions [4, 5] is feasible and has shown value. For example, by analyzing human-created artifacts and interactions, we can create a model of individual expertise or communities of interest [7]. In information seeking sessions, detecting selections and rejections of information provides an opportunity to induce individual and group profiles that can assist in content selection and presentation generation.

Persistence/transience of interest profiles. User information needs tend to change over time, with profiles rapidly becoming out of date. Monitoring user queries and story selections is one method that can address this problem. Generalizing from users' specific interests can yield an even richer user model.

Tailoring. More sophisticated mechanisms are required to tailor content to specific topics or users. In addition to content selection, material must be ordered and customized to individual user interests, platforms, or task/cognitive situations. This will require methods of presentation generation that integrate extracted or canned text with generated text.

Information extraction. Over the longer term we are working to create techniques to automatically summarize, fuse, and tailor selected events and stories. This requires deeper understanding of the source news material beyond extracting named entities, key frames, or key sentences.

Multilingual content. Because news is global in production and dissemination, it is important to support access to and integration of foreign language content. This poses not only multilingual processing challenges but also requires dealing with different country/cultural structures and formats.

Cross source/story fusion. Another important problem is not only summarization of individual stories but summarizing across many stories, possibly from different sources or languages. This is particularly challenging when the sources may be inconsistent in purpose, content, or form.

Evaluation. Community-defined multimedia evaluations will be essential for progress. Key to this progress will be a shared infrastructure of benchmark tasks with training and test sets to support cross-site performance comparisons.

Conclusion

This article envisions user-tailored question answering from multimedia sources. We've demonstrated how user performance and enjoyment can be dramatically increased by a mixture of multimedia extraction and summarization, user modeling, and presentation planning tailored to specific user interests and preferences. We've also outlined a range of research frontiers that promise new challenges to the question answering research community. C

References

1. Advanced Question Answering for Intelligence (AQUAINT). Center for Intelligent Information Retrieval; www.ic-arda.org/InfoExploit/aquaint.
2. Boykin, S. and Merlino, M. A. Machine learning of event segmentation for news on demand. *Commun. ACM 43*, 2 (Feb. 2000), 35–41.
3. Breck, E., Burger, J. D., Ferro, L., House, D., Light, M., and Mani, I. A sys called Qanda. In E. Vorhees and D. Harman, Eds. *The Eighth Text Retrieval Conference*, NIST Special Publication, Feb. 2000, 499–506.
4. Linton, F., Joy, D., Schaefer, H-P., and Charron, A. *OWL: A Recommender System for Organization-Wide Learning.* Educational Technology and Society, 2000; ifets.ieee.org/periodical/.
5. Linton, F., Joy, D., and Schaefer, H-P. Building user and expert models by long-term observation of application usage. In J. Kay, Ed., *UM99: User Modeling: Proceedings of the Seventh International Conference.* Springer Verlag, New York, 1999, 129–138; selected data accessible from an archive on zeus.gmd.de/ml4um/.
6. Maybury, M. News on demand: Introduction. *Commun. ACM 43*, 2 (Feb. 2000), 32–34.
7. Maybury, M., D'Amore, R, and House, D. Expert finding for collaborative virtual environments. *Commun. ACM 44*, 12 (Dec. 2001).
8. Maybury, M., Merlino, A., and Morey, D. Broadcast news navigation using story segments. *ACM International Multimedia Conference* (Seattle, WA, Nov. 8–14, 1997), 381–391.
9. Merlino, A. and Maybury, M. An empirical study of the optimal presentation of multimedia summaries of broadcast news. I. Mani and M. Maybury, Eds., *Automated Text Summarization*, MIT Press, 1999, 391–402.
10. Voorhees, E. and Tice, D.M. The TREC-8 question answering track evaluation. In E. Voorhees and D.K. Harman, Eds., *The Eighth Text Retrieval Conference (TREC-8)*, NIST Special Publication 500-246, 2000, 83–106.

Marc Light (light@mitre.org) is a principal scientist in the Information Technology Division at the MITRE Corporation, Bedford, MA.

Mark T. Maybury (maybury@mitre.org) is the executive director of the Information Technology Division at the MITRE Corporation, Bedford, MA.