Appears in IEEE Advances in Digital Libraries, Library of Congress, Washington, DC, May 2000

Audio Structuring and Personalized Retrieval Using Ontologies

Latifur Khan and Dennis McLeod
Department of Computer Science and
Integrated Media Systems Center
University of Southern California
Los Angeles, California 90089
[latifurk, mcleod]@usc.edu

Abstract

The goal of this work is to improve the accuracy (precision and recall) and communication effectiveness of a database system response to a user information request, by utilizing a domain-specific ontology. This ontology is employed, along with user profile information, to automatically select and deliver appropriate information units from a database. Specifically, the focus here is on multimedia audio databases: we are developing a Personal AudioCast system at the USC Integrated Media Systems Center which creates and delivers to an end-user personalized audio programs constructed from an audio database. Techniques are developed to segment and structure audio information, process user requests to generate custom audio results, create and optimize SQL queries, and present the results to the user.

1. Introduction

The field of digital media continues to be heavily impacted by technical advances. These advances are changing the nature of the information generated by these media. A great deal of textual information is now being augmented by ever increasing amounts of non-textual information: images, video, and audio. New information sources are also being created by technical advances which allow the production of purely non-textual information. One can expect the future to bring continued growth in this explosion of information and its changing nature in the direction of multimedia. In fact, it is clearly the case that in the area of non-textual information technological development is still in its infancy. Effective solutions for the management of the growing quantum of multimedia information will depend upon the maturation of several important new technologies [16].

Audio is one of the most powerful and expressive non-textual media. Moreover, audio information can be of significant benefit to a visually impaired person. Audio is a streaming medium (temporally extended), and its properties make it a popular medium for capturing and presenting information. At the same time, these very properties, along with audio's opaque relationship to the computers, present several technical challenges from the data management perspective. One of the key challenges is finding ways to facilitate the retrieval of audio information through the provision of descriptions (*metadata*). Effective selection/retrieval of audio entails several tasks, such as metadata generation, management of metadata, and selection of audio information. In this paper, We have devised general-purpose, scalable techniques and mechanisms for user-customized selection and presentation of multimedia information. Specifically, we have proposed a potentially powerful and novel approach for the customized selection/retrieval of multimedia information. The crux of our innovation is the development of an ontology-based model that facilitates to achieve minimal irrelevant information (high precision) and minimal loss of information (high recall). An ontology is a collection of key concepts and their inter-relationships that collectively provide an abstract view of an application domain.

Historically, ontologies have been employed in text retrieval systems in order to achieve high precision and high recall. Attempts have been made to deal with the problem of query expansion with semantically related-terms [19, 29], and to carry out the comparison of queries and documents via conceptual distance measures [30, 21]. The use of ontology for query expansion and the disambiguation of document terms using ontology are manual procedures which are both subjective and labor intensive. Further, only Voorchees [29] has reported results which are promising for short and

incomplete queries. Our proposed ontology-based model expands queries automatically, which enhances recall. Further, we present a new automatic disambiguation mechanism for metadata with ontology that will also improve precision. At present, an experimental prototype for the implementation of the model is in development. As of today, our working ontology has around 6,500 concepts for the sports news domain, with 12 hours of audio stored in the database.

To demonstrate the power of our model we have obtained an operational result which indicates the superiority of ontology-based search techniques over keyword-based methods. For this illustration we have chosen sample queries which are related to the most general and the most specific descriptions. We have observed that in the case of the most general descriptions ontology-based model out performs keyword-based technique in terms of both precision and recall.

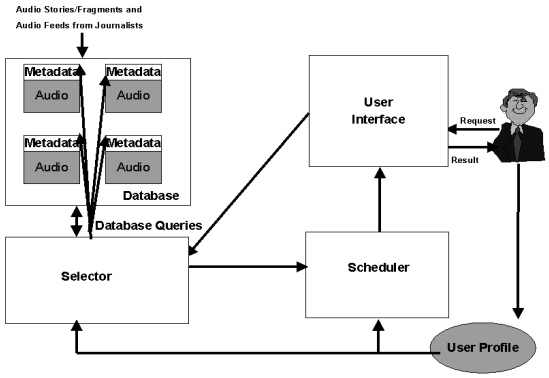

Figure 1. PAC Functional Architecture

2. Background

Figure 1 presents the functional architecture of a user-customized audio information-on-demand system which we term Personal AudioCast (PAC). A user request is processed by the user interface and dispatched to the PAC selector. The selector then constructs a script for a user personalized audio package, employing the database and a profile for the user. Selections from the PAC database are assembled and packaged for the user as a continuous audio presentation. The selector then passes the script to the PAC scheduler which dispatches and delivers the package to the user. For example, several hours of broadcast sports audio are stored in a PAC database along with descriptions of audio (metadata). Let us assume a hypothetical user who is only interested in hockey and basketball. These interests are included in the profile for that particular user. Now, suppose this user sends a request by specifying "Give me sports news." The user Interface handles this request and sends it to the PAC selector. Consulting the user profile, the PAC selector generates database queries related to hockey and basketball, and submits these to the database. Relevant audio news items are identified and the scheduler starts to play them. Thus, by employing a user profile, unwanted news items are automatically filtered out.

Given the above framework for the PAC system, we can identify key research and engineering problems/tasks that must be addressed:

- Segmentation: Audio (broadcast) may consist of multiple news items. In general, some items are of interest to the user and some are not. Therefore, we need to identify the boundaries of news items of interest so that the user's query can directly retrieve these segments. A mechanism is required which will make this possible. A change of speaker and long pauses are used to identify segment boundaries [4]. Further, multiple contiguous segments may form a news item.
- Metadata acquisition: To retrieve particular news items we need to provide descriptions that will be used to match user requests to the information provided. These descriptions along with boundaries of news items, are termed *metadata*. Metadata, along with audio, are stored in the database for retrieval. The accuracy of retrieval depends on the technique employed for the generation of metadata. Following current state-of-the-art in speech recognition technology, we employ a semi-automatic technique for the generation of metadata using word-spotting [24] (which determines occurrence of keywords in audio). The technique employed in the generation of metadata utilizes an ontology (see Figure 2). In this paper, we use a domain dependent ontology in order to facilitate metadata generation involving an arbitrary number of contiguous segments. Further, we propose the effective management of metadata in the database by considering metadata in terms of specific concepts, rather than generalized concepts from the ontology.
- Selection: A mechanism is required which converts user requests into database queries used to retrieve relevant audio. These queries are then executed and the results sent to the scheduler. The mechanism should be sufficiently robust to efficiently support different types of queries. We demonstrate that

customized audio selection can be provided by the ontology-based model. For selection, user explicit requests and user profiles are first associated with the concepts of the ontology. Specific concepts for each of these associated concepts are then generated from the ontology and expressed in disjunctive form for database query. It is important to note that a user profile is simply a collection of concepts. For database queries, we use the most widely used query language, SQL.

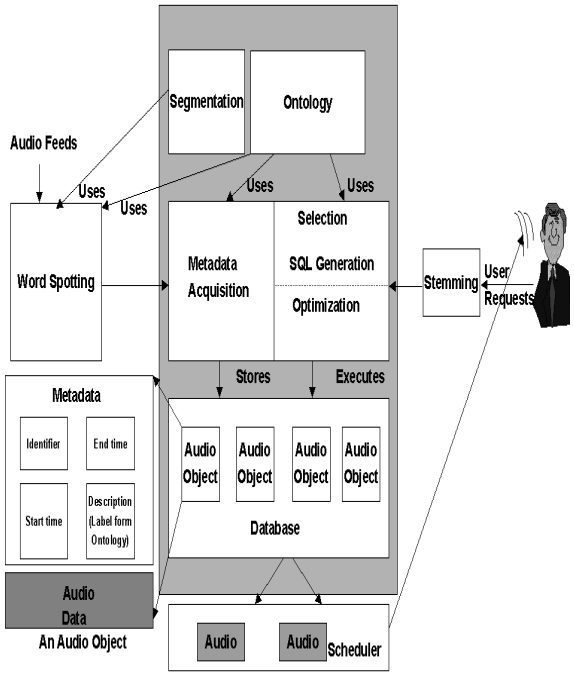

Figure 2. Architecture for Audio Customization

- Scheduling: Selected audio segments should be played with minimal delay by utilizing streaming and caching. For scheduling we rely on a commercial product, Java media framework, which facilitates streaming and caching during presentation of audio.
- User profile generation: To support personalization, a PAC user profile specifies characteristics, preferences, and communication style for tailoring a periodic or on-demand delivery to the user. Thus, user profile information is obtained by monitoring the user, as well as through explicit user input. For user profile generation the user may specify concepts by browsing the ontology through a user interface. Using these concepts from a user profile, the selector generates SQL queries whenever a user accesses the PAC. The scheduler then fetches relevant information, rather than retrieving the entire audio content. To enable user profiles to evolve toward true user preferences, such as those made possible by learning, we assume that an agent [18] can be employed to capture user implicit preferences during the presentation of audio.

In this paper, we focus on the key research issues of segmentation, metadata acquisition, and selection (these are shown as the rectangular box with shading in Figure 2). The remainder of this paper is organized as follows. Section 3 covers some related work on customization, metadata acquisition, and selection. Section 4 describes a technique for the segmentation of audio information. Section 5 describes how we can use an ontology which will support metadata generation for audio, and discusses different mechanisms for annotating audio based on current state-of-the-art speech technology. Section 6 demonstrates how this ontology can be used to generate information selection requests in SQL query. Section 7 describes the current implementation status of PAC. Finally, section 8 presents our conclusions and future work.

3. Related Work

To place our research in the context of personalization and multimedia domains, we summarize key related efforts. First, we discuss different customization techniques to achieve personalization. Second, we present different efforts for metadata acquisition. Third, we present a number of different techniques using ontologies in a text retrieval system to improve precision and recall. Finally, we present key information modeling techniques for the selection of multimedia information.

In general, customized selection and presentation are employed to bridge the gap between users and the providers of information in a variety of application contexts. Customized selection, in this case, involves the process of tailoring available document contents in the information repository to satisfy a set of user preferences, and to reflect long-term user interests in an efficient way. There are a number of active efforts under way to develop systems for customized selection and presentation. The main strands of related research for web are CRAYON [27], Fishwrap [28], and Krakatoa Chronicle [26]. Among these, the Krakatoa Chronicle is the most recent. It is an experimental system which implements an interactive, personalized newspaper on the WWW.

For metadata acquisition we need to specify the content of media objects. Two main approaches have been employed to this end: fully automated content extraction [10] and selected content extraction [24]. Due to the weakness of the fully automated content extraction in state-of-art audio/speech recognition, we have chosen

the latter approach (which is semi-automatic). This is because as Hauptman has shown in the Informedia project that automatic transcription (indexing) of speech is a difficult task. This results from the fact that the current speech recognition systems support limited vocabulary, at least an order of magnitude smaller than that of a text retrieval system [25]. Additionally, environmental noise generates inevitable speech recognition error. Therefore, to insure appropriate selection and presentation of audio information, we need to capture its semantic description. By using word-spotting, for selected content extraction speech recognition looks for a set of predefined keywords in audio where these keywords should convey semantic descriptions. These keywords are from an ontology. For example, the video mail retrieval group [15] has investigated using 35 pre-selected keywords for audio information retrieval and reported near 90% retrieval accuracy.

Historically, ontolgoies have been employed in text retrieval systems in order to achieve high precision and high recall. Attempts have been made in two directions. At the one end is query expansion with semantically related-terms and at the other end is the comparison of queries and documents via conceptual distance measures. Query expansion with a generic ontology, WordNet [19] has shown to be potentially relevant to enhance recall, as it permits matching relevant documents that could not contain any of the query terms. Voorchees [29] manually expands 50 queries over a TREC-1 collection using WordNet and observes that expansion was useful for short, incomplete queries. But expansion was not promising for complete topic statements. Further, for short queries automatic expansion is not trivial; it may degrade retrieval performance rather than enhancing it. The notion of conceptual distance between query and document provides an approach to modeling relevance. Smeaton et. al. [30] and Gonzalo et. al. [21] focus on managing short and long documents, respectively. Note here that queries and document terms are manually disambiguated using WordNet. This manual technique is subjective and labor intensive.

Although we use audio, here we show related work in the video domain which is closest to and which complements our approach in the context of data modeling for the facilitation of information selection requests. Key related work in the video domain for selection of video segments includes [1, 2, 11]. Of these, Omoto et al. use a knowledge hierarchy to facilitate annotation, while others use simple keyword based techniques without a hierarchy. The model of Omoto et al. fails to provide a mechanism that automatically converts a generalized description specialized one(s). Further, this annotation is manual and it does not deal with the disambiguation issues related to concepts.

4. Segmentation of Audio

Since audio is by nature serial totally, random access to audio information may be of limited use. To facilitate access to useful segments of audio information within an audio recording, we need to identify entry points/jump locations. Either a change of speaker or long pauses can serve to identify entry points [Aron94]. However, the drawback of this approach is that it may fail when the speaker starts with the news items in a same breath. For long pause detection, we use short time energy (E_n). The short time energy (E_n) provides a measurement for distinguishing speech from silence [22] for a frame (consisting of fixed number of samples) which can be calculated by the following equation:

$$E_n = \Sigma_{m=-\infty \text{ to } \infty} [\, x(m)\, w(n-m)\,]^2$$
$$= \Sigma_{m=n-N+1 \text{ to } n}\ x^2(m)$$

Where x(m) is discrete audio signals, n is the index of the short-time energy, and w(m) is a rectangle window of length N. When the E_n falls below a certain threshold, we treat this frame as pause. After detection of pause, in terms of frame, we combine several adjacent pauses and generate long pauses. By employing the above strategy, long pauses are identified. Therefore, the presence of speeches with starting and ending points allow us to detect the boundaries of audio segments.

An audio object is thus composed of a sequence of contiguous segments. In our model, however, the start time of the first segment and the end time of the last segment of these contiguous segments are used to denote start time and end time of the audio object, respectively. Thus, in our model, pauses between interior segments are kept intact to provide intelligibility of speech. The formal definition of an audio object indicates that an audio object's description is provided by a set of self-explanatory tags or labels.

Formal Definition of an Audio Object:

An audio-object O_i is defined by five tuple (Id_i, S_i, E_i, V_i, A_i) where

- Identifier: Id_i is an object identifier which is unique
- Start time: S_i is the start time
- End time: E_i is the end time. Start time and end time satisfy $E_i - S_i > 0$
- Description: V_i is a finite set of tags or labels, i.e., $V_i = \{v_{1i}, v_{2i}, \ldots, v_{ji}, \ldots, v_{ni}\}$ for a particular j where v_{ji} is a tag or label.
- Audio data: A_i is simply audio recording for that time period.

For example, an audio object is defined as {10, 1145.59, 1356.00, {Wayne Gretzky}, *}. Here, the

identifier of the object is 10, start time and end time are 1145.59, and 1356.00 unit respectively, and description is "Wayne Gretzky", and * denotes audio data.

Of the information in the five tuple, the first four items (identifier, start time, end time, and description) are called *metadata*.

5. Metadata Acquisition

In this section, we first describe how we can use ontology to facilitate metadata generation. Second, we present different annotation techniques. Third, we present a disambiguation mechanism with reference to selected concepts for a group of keywords. Finally, we present a discussion of metadata management issues raised here.

An ontology is a specification of an abstract, simplified view of the world that we wish to represent for some purpose [5, 9]. Therefore, an ontology defines a set of representational terms that we call *concepts.* Inter-relationships among these concepts describe a target world. An ontology can be constructed in two ways, domain dependent and generic. CYC [17], WordNet [19], or Sensus [20] are examples of generic ontologies. For our purposes, we choose a domain dependent ontology. First, this is because a domain dependent ontology provides concepts in a fine grain, while generic ontologies provide concepts in coarser grain. Second, a generic ontology provides a large number of concepts that may contribute large speech recognition error.

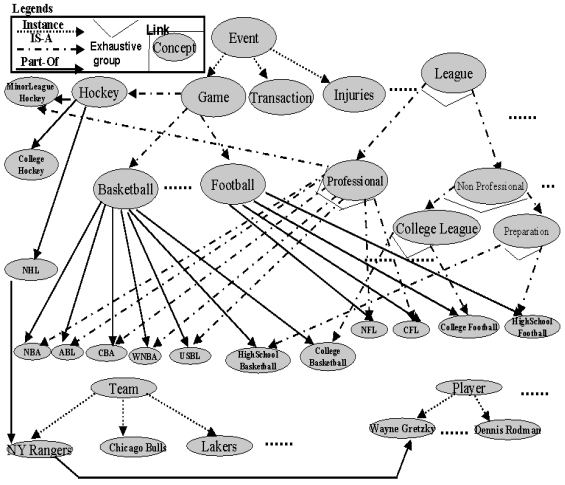

Figure 3. A Portion of an Ontology for the Sports Domain

Figure 3 shows an example ontology for sports news. This ontology is usually obtained from generic sports terminology and domain experts [6]. This ontology is described by a directed acyclic graph (DAG). Here, each node in the DAG represents a concept. In general, each concept in the ontology contains a label name and a synonyms list. Note also that this label name is unique in the ontology. Further, this label name is used to serve as association of concepts with audio objects. The synonyms list of a concept contains vocabulary (a set of keywords) through which the concept can be matched with user requests. Note that a keyword may be shared by multiple concepts' synonymous lists. For example, player "Bryant Kobe," "Bryant Mark," "Reeves Bryant" share common word "Bryant" which may create ambiguity problem. Moreover, each of them belongs to league NBA. Hence, each of these concepts' label is prefixed with concept, NBA's label that allows to make efficient query generation for upper level concepts (see Sec. 6).

5.1. Inter-Relationships

In the ontology, concepts are interconnected by means of inter-relationships. If there is a inter-relationship R, between concepts C_i and C_j, then there is also a inter-relationship R′ between concepts C_j and C_i. In Figure 3, inter-relationships are represented by labeled arcs/links. Three kinds of inter-relationships are used to create our ontology: IS-A, Instance-Of, and Part-Of. These correspond to key abstraction primitives in object-based and semantic data models [3].

IS-A: This inter-relationship is used to represent concept inclusion. A concept represented by C_j is said to be a specialization of the concept represented by C_i if C_j is kind of C_i . For example, "NFL" is a kind of "Professional" league. In other words, "Professional" league is the generalization of "NFL." In Figure 3, the IS-A inter-relationship between C_i and C_j goes from generic concept C_i to specific concept, C_j represented by a broken line. The IS-A inter-relationship can be further categorized into two types: *exhaustive group* and *non-exhaustive group*. An exhaustive group consists of a number of IS-A inter-relationships between a generalized concept and a set of specialized concepts, and places the generalized concept into a categorical relation with a set of specialized concepts in such a way so that the union of these specialized concepts is equal to the generalized concept. For example, "Professional" relates to a set of concepts, "NBA", "ABL", "CBA", ..., by exhaustive group (denoted by caps in Figure 3). Further, when a generalized concept is associated with a set of specific concepts by only IS-A inter-relationships that fall into the exhaustive group, then this generalized concept will not participate in the annotation and SQL query generation explicitly. This is because this generalized concept is entirely partitioned into its specialized concepts through an exhaustive group. We

call this generalized concept a *non participant concept (NPC)*. For example, in Figure 3 "Professional" concept is NPC. On the other hand, a non-exhaustive group consisting of a set of IS-A does not exhaustively categorize a generalized concept into a set of specialized concepts. In other words, the union of specialized concepts is not equal to the generalized concept.

Specialized concepts inherit all the properties of the more generic concept and add at least one property that distinguishes them from their generalizations. For example, "NBA" inherits the properties of its generalization, "Professional" but is distinguished from other leagues by the type of game, skill of participant, and so on.

Instance-Of: This is used to show membership. A C_j is a member of concept C_i. Then the inter-relationship between them corresponds to an Instance-Of denoted by a dotted line. Player, "Wayne Gretzky" is an instance of a concept, "Player." In general, all players and teams are instances of the concepts, "Player" and "Team" respectively.

Part-Of: A concept is represented by C_j is Part-Of a concept represented by C_i if C_i has a C_j (as a part) or C_j is a part of C_i. For example, the concept "NFL" is Part-Of "Football" concept and player, "Wayne Gretzky" is Part-Of "NY Rangers" concept.

When a number of concepts are associated with a parent concept through IS-A inter-relationship, it is important to note that these concepts are disjoint, and are referred to as concepts of a disjoint type. When, for example, the concepts "NBA", "CBA", or "NFL" are associated with the parent concept "Professional," through IS-A, they become disjoint concepts. Moreover, any given object's metadata cannot possess more than one such concept of the disjoint type. For example, when an object's metadata is the concept "NBA," it cannot be associated with another disjoint concept, such as "NFL." It is of note that the property of being disjoint helps to disambiguate concepts for keywords during annotation.

Concepts are not disjoint, on the other hand, when they are associated with a parent concept through Instance-Of or Part-Of. In this case, some of these concepts may serve simultaneously as metadata for an audio object. An example would be the case in which the metadata of an audio object are team "NY Ranger" and player "Wayne Gretzky," where "Wayne Grezky" is Part-Of "NY Rangers."

5.2. Annotation

Annotation is the name for the process through which concepts are associated with audio objects. The completeness and accuracy of the annotation process contributes to the success of customization. In this section, we describe techniques for automatic and manual annotation.

Automatic Annotation: Word-spotting techniques can provide the selected content extraction to make the annotation process automatic. Word-spotting, as noted, is a particular application of automatic speech recognition in which the vocabulary of interest is relatively small. Vocabularies of concepts from the ontology, excepting NPC concepts, can be used in our case. It is the job of the recognizer to pick out only occurrences of keywords from this vocabulary in the speech to be recognized [14]. The output of a wordspotter is typically a list of keyword "hits," labeled with each keyword's start timing. From these timings, we can determine what type of information audio segments convey. Note that automatic annotation obviously results in a great reduction in human labor and a substantial increase in scalability.

After applying segmentation and word-spotting techniques, automatic annotation can be done in the following way. When new concept(s) are detected in a segment, a new object is created. The start time of this new object is equal to the current segment's start time, and metadata is provided by the label name of this concept. In addition, the end time of the new object is kept to the current segment's end time. It may later be updated. This is because recall that an audio object may be composed of a number of contiguous segments. On the other hand, if a set of successive segments does not contain any new segment, or continues to contain the same concept, then the already defined object in an earlier segment will cover this set of segments until another new concept is recognized in the next segments. In other words, the end time of this existing object will be updated to the end time of the last segment from this set.

For example, if the occurrence of the concept, "NFL" is determined in a particular segment, an object is defined and its metadata is the concept, "NFL." If a set of successive segments does not contain a new concept, we can simply assume that this object will cover this next set of segments until a different concept is recognized in a new segment. In other words, this object's end time will be updated to the last segment's end time of this set of successive segments in which no new concept has occurred. It is important to note that a keyword may be recognized for which multiple concepts will be selected due to partial match with concepts' synonyms list.

Manual Annotation: Due to incompleteness of ontology, an audio object may not be annotated with any concept. In this case, manual annotation will be necessary. In manual annotation arbitrary sequences of contiguous segments are described rather than describing each segment independently, which was strongly criticized by Smith et al [23], due to its inflexibility.

Note that these sequences of contiguous segments form an audio object which is consistent with our audio object definition. The process of creating relationships between objects and descriptions (here concepts) is known as *stratification*.

5.3. Disambiguation

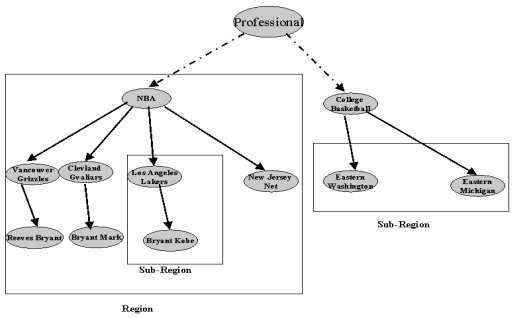

Figure 4. Disambiguation of Concepts

False matches, with a consequent loss of precision, can result from the association of a single keyword with more than one concept. In our ontology, for example, "Charlotte Hornets" and "UNC Charlotte" are two teams under "NBA" and "College Basketball," respectively. Thus, the keyword provided by word-spotting or manual annotation, "Charlotte" is associated with two concepts. In practice, we require a procedure that can resolve which concept is intended from the use of this keyword.

The disambiguation technique used to accomplish this task is based on the idea that a set of keywords occurring together in context will determine appropriate concepts for one another, in spite of the fact that each individual keyword is multiply ambiguous. For example, a set of nouns, such as base, bat, glove, and hit may refer to several concepts when each is taken alone. But, when taken together, the intent is clearly a reference to the concept of baseball.

For automatic disambiguation a set of *region*s representing different concepts needs to be defined. The keywords in a given annotation will then fall into a number of regions representing various concepts.

For disambiguation of concepts, first, we determine a region where the number of selected concepts appeared is maximum. Note that regions consist of disjoint concepts, so regions are mutually exclusive. Thus, the region with maximum number of selected concepts will be used for annotation; other selected concepts of different regions are automatically pruned. Next, we may apply the above pruning technique recursively until no more ambiguity exists.

Obviously the manner in which regions are defined will be a critical component of this procedure. Formally, a region is a tree extracted from our ontology (represented by DAG), where the root is a disjoint type concept which can help define the boundary of each region. For example, NBA and CBA, two different associations of professional, are disjoint concepts. Each of these concepts and its children concepts can be used to define a region. Thus NBA, and its teams and players, form a region. CBA and its teams and players will form a different region.

In a given annotated text, concepts are first selected based on matching of each concept's synonyms list with the keywords of the annotated text. Several important points need to be noted in connection with this procedure.

First, matching between the keywords of the annotated text and the synonyms lists related to concepts may be partial or incomplete. This requires that each concept selected be associated with a normalized maximum score (weight) based on the number of words in each element of the synonyms list which have been matched with words in the annotated text. Thus some concepts selected will have higher scores than others. In this case, for each region, we need to add the scores of selected concepts and choose a region with the highest score.

It is important to note further that the identification of a single region may not be adequate to disambiguate keywords. This is because the same keyword may still select more than one concept in this region. Thus, within this region we will be required to apply the same scoring technique recursively, identifying sub-regions which correctly disambiguate words. Each sub-region will consist of a number of concepts directly connected via one of inter-relationship link or a single concept.

Suppose, for example, that the annotated text for a particular object is "*Kobe. Lakers keep grooving with 8th straight win. Kobe Bryant scores 21 points as the lakers remain perfect on their eastern road trip with a 97-89 triumph over the Nets. Bryant discussed the eight game win streak and his performance in the All Star game."*

The concepts selected by this match are "Los Angeles Lakers," "New Jersey Nets," "Bryant, Kobe," "Bryant Mark," "Reeves Bryant," "Eastern Washington" and "Eastern Michigan" (see Figure 4).

With the exception of Eastern Washington and Eastern Michigan, which are part of the concept college basketball, the concepts belong to a region in which the root concept is "NBA." Then, the number of concepts appearing in the NBA region is maximum. Note that several players are selected for keyword "Bryant." However only "Kobe Bryant" is part of team "Los

Angeles Lakers" (inter-relationship) which is already selected. Thus, a sub-region consisting of concepts "Bryant Kobe" and "Los Angeles Lakers" is selected. Other concepts selected for keyword "Bryant" in this region are pruned.

5.4. Metadata Management

Effective management of metadata facilitates efficient storing and retrieval of audio information. To this end, in our model most specific concepts are considered as metadata. Several concepts of the ontology, for example, can become the candidate for the metadata of an audio object, regardless of the annotation technique employed. However, some of these may be children of others. Several approaches can be used to address this problem.

First, we can simply store the most general concepts. But we may get many irrelevant objects (precision will be hurt) for queries related to specific concepts. Second, the most specific concepts can be stored in the database. In this case, in order to support queries related to the most general concepts all the concepts associated with these objects are stored, thus facilitating efficient retrieval of objects for such concepts.

There are drawbacks to this second approach. First, a huge amount of storage is required. Along with its specific concepts, all its parent concepts must be in the database. For example, for player, "Bryant Kobe" a team name, league name, and all other parent concepts must be stored along with the player concept,"Bryant Kobe." Further, the inter-relationships of concepts in an ontology may change, requiring update operations in the database.

An alternative solution is to deploy the most specific concepts as metadata. Corresponding generalized concepts can then be discarded. Suppose, for example, an audio object becomes the candidate for the concepts "NHL", "Hockey", and "Professional." During the annotation process the object will only be annotated with the most specific concept, "NHL." In this case, the metadata of the audio objects stored in the database will be comprised of the most specific concepts.

By storing specific concepts as metadata, rather than generalized concepts of the ontology, we can expect to achieve the effective management of metadata. Even so, the audio object, in the above example, can still be retrieved through querying the system by "NHL", "Hockey", and "Professional." This is because user requests are first passed through ontology on the fly and expressed in terms of most specific concepts.

Here, we consider an efficient way of storing audio objects in the database: we maintain a single copy of all the audio data in the database. Further, each object's metadata are stored in the database. An *interval* of an audio object is defined by start time and end time. Thus, this start time, and end time of an object point to a fraction of all the audio data. Therefore, when the object is selected, this boundary information provides relevant audio data that are to be fetched from all the audio data and played by the scheduler. The following self-explanatory schemas are used to store audio objects in the database: *Audio_News (Id, Time_Start, Time_End, ...), and Meta_News (Id, Label)*. Each audio object's start time, end time and description correspond to Time_Start, Time_End, and Label respectively. Thus, each object's description is stored as a set of rows or tuples in the Meta_News table for normalization purpose.

6. Selection

We now focus specifically on our techniques for utilizing an ontology-based model for processing information selection requests. Here, we discuss a technique for SQL query generation.

In response to a user request, the basic intuition for the generation of a SQL query is as follows. We assume that user requests are expressed in plain English. Tokens are generated from the text of the user's request after stemming. These tokens are associated with concepts in the ontology using synonyms list. We call each of these associated concepts a *QConcept*. Among these QConcepts, some concepts are the children of other concepts in the ontology. In this case, we discard parent concepts. Recall that specific concepts are annotated with audio objects.

Let us now consider how each QConcept is mapped into the "where" clause of SQL query. Note that by putting the QConcept as a boolean condition in the "where" clause, we are able to retrieve relevant audio objects. First, the QConcept is located in the DAG of the ontology by matching the user's request to the concept's synonyms list. Next, we check whether the concept is a NPC type. If it is already a NPC concept, it will not be used as a boolean condition in the "where" clause of SQL query. Recall that NPC concepts can be expressed exhaustively as a collection of more specific concepts. If the concept is not of the NPC type, its label name is added in the "where" clause. If the concept is not leaf node, all of its children concepts are generated using depth/breadth first search (DFS/BFS). For each of them, the concept's category is checked. If it is not NPC, the concept's label name will be added to the "where" clause. Further, if it is not a leaf concept, its children concept will be generated and the above technique will be recursively applied. One important observation is that all concepts appearing in an SQL query for a particular QConcept are expressed in disjunctive form. Note that

we assume SQL query is generated at the client site, submitted over the network to the server, and executed at the server.

The following example illustrates the above process. Suppose the user request is "Please give me news about player Bryant Kobe." "Bryant Kobe" turns out to be the QConcept which is itself a leaf concept. Hence, the generated SQL query by using only "Bryant Kobe" (label is "NBAPlayer9") is:

SELECT Time_Start, Time_End
FROM Audio_News a, Meta_News m
WHERE a.Id=m.id
AND Label="NBAPlayer9"

Let us now consider the user request, "Tell me about Los Angels Lakers." Note that the concept "Los Angeles Lakers" is not NPC, so its label ("NBATeam11") will be added in the "where" clause of SQL query. Further, this concept has several children concepts ("Bryant Kobe," "Celestand John," "Horry Robert," ....) names of players for this team. Note that these player concepts' labels are "NBAPlayer9," "NBAPlayer10," and "NBAPlayer11," respectively. In SQL query:

SELECT Time_Start, Time_End
FROM Audio_News a, Meta_news m
WHERE a.Id = m.Id
AND (Label="NBATeam11"
OR Label="NBAPlayer9"
OR Label="NBAPlayer10"...)

Since most specific concepts are used as metadata and our ontologies are large, in the case of querying upper level concepts, every relevant child concept will be mapped into the "where" clause of the SQL query and expressed as a disjunctive form. To avoid the explosion of boolean conditions in this clause of the SQL query, the labels for the player and team concepts are chosen in a tricky way. These labels begin with the label of the league in which the concepts belong. For example, team "Los Angeles Lakers" and player "Bryant, Kobe" are under "NBA." Thus, the labels for these two concepts are "NBATeam11" and "NBAPlayer9" respectively, whereas the label for the concept "NBA" is "NBA."

Now, when user requests come in terms of an upper level concept (e.g. "Please tell me about NBA.") the SQL query generation mechanism will take advantage of prefixing:

SELECT Time_Start, Time_End
FROM Audio_News a, Meta_News m
WHERE a.Id=m.Id
AND Label Like "%NBA%"

7. Experimental Implementation

The development of the PAC system is still in progress. However, from its current state we can report on the foundation of the choices that have been initially taken by our project. The system is based upon a client server architecture: the server (a SUN Sparc Ultra 2 model with 188 MBytes of main memory) has an Informix Universal Server (IUS) [13], which is an object relational database system. For the sample audio content we use CNN broadcast sports audio [7] and Fox Sports. We have written a hunter program which goes to these web sites and downloads all audio and video clips with closed captions. On average, each of the clips is 1.5 minutes in length. The average size of the closed captions is 112 words. As of today, the total duration of stored audio is 12 hours. Wav and ram are used for media format.

To detect the boundary of new items we need to detect long pauses and speaker changes. In the current implementation, we have considered the only pause detection by considering the energy function of signals. We chose frame size 100 milliseconds. We determined maximum and minimum average energy value of the frames for a particular day's recorded audio. From these two, we fixed a threshold (T) point for that day that is equal to minimum average energy plus 20% of the difference of maximum and minimum average energy. In our case, we observed that this particular value of T worked well to detect pause after doing several experiments on the selection of values of T. Next, we merged adjacent pauses together and determined a long pause by setting a pause threshold. This threshold value was also determined in a similar way to T by considering maximum and minimum pause length.

Figure 5 demonstrates the current query interface and media player which was written in Java using Java Media Framework API. Through this interface, the user submits queries in plain English; the selection mechanism generates SQL query which is sent to the database system. However, for conjunctive or difference query, users express keywords with boolean operators. After that, the scheduler module receives a set of intervals as a result of the execution of the SQL query, and next, the Scheduler fetches corresponding audio data, using the HTTP protocol. In Figure 5 "Next Object", "Previous Object", and "Repeat Object" buttons facilitate to listen next, previous, and current audio object respectively. In addition, this interface shows how many audio objects have been selected for a particular query, and what percentage of the current object that is played so far, and so on.

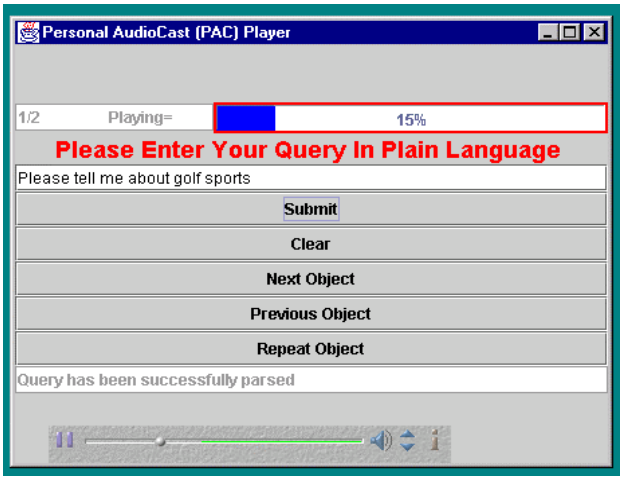

Figure 5. Query Interface and Player for PAC

Currently, our working ontology has around 6,500 concepts for sports domain. For fast retrieval, we fetch the upper level concepts of the ontology in main memory and leaf concepts are fetched on a demand basis. Hashing is also used to increase the speed of retrieval. As of today, 480 audio objects have been defined and their associated closed captions are used to hook with the ontology.

Results have been improved by employing stemming, and by eliminating stop words from the closed captions. During the automatic annotation of concepts with audio objects we have observed that 84% of the objects are correctly annotated and 16% of the objects are missed due to the incompleteness of the ontology. Further, 8% of the objects have been annotated with wrong concepts along with the concepts which are correct.

An illustration of one result we have obtained which demonstrates the power of ontology based over keyword based technique involves the consideration of some sample queries which are related to upper level concepts. We have observed that in these queries the ontology-based model out performs keyword based techniques in terms of both precision and recall (see Figure 6).

In Figure 6 the first two bars represent comparative results for precision. For an upper level concept we use a query like "Please tell me about the NBA." In this case, correct results for recall and precision for the ontology-based model are 98% and 97% respectively. On the other hand, results for recall and precision for the keyword based technique are 16% and 90% respectively. This is because the keyword based technique can only retrieve objects where the keyword NBA is present, reducing the amount of recall. Ontology helps expand queries which contain semantically similar terms, thus recall will be improved. Further, during annotation the disambiguation mechanism in the ontology-based model helps to achieve better precision by keeping only NBA and throwing out other words from annotated text which may serve as a match for a keyword-based model. For leaf concepts, ontology cannot expand queries with semantically similar terms. Thus, recall is the same for both cases.

8. Conclusions

In this paper we have proposed a potentially powerful and novel approach for the retrieval of audio information. The crux of our innovation is the development of an ontology-based model for the generation of metadata for audio, and the selection of audio information in a user customized manner. We have shown how the ontology we propose can be used to generate information selection requests in database queries. We have used a domain of sports news information for a demonstration project, but our results can be generalized to fit many important content domains including but not limited to all audio news media. Our ontology-based model demonstrates its power over keyword based search techniques by providing many different levels of abstraction in a flexible manner.

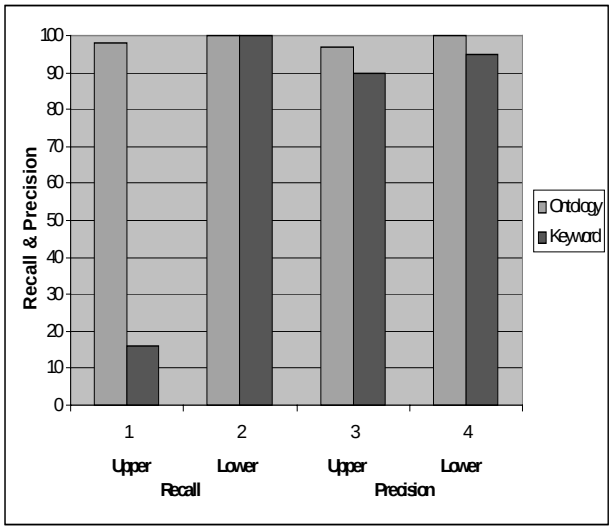

Figure 6. Power of Ontology-Based over Keyword-Based Search Techniques

Although we are confident that the fundamental conceptual framework for this project is sound, and its implementation completely feasible from a technical standpoint, some questions remain to be answered in future work. These include detailed work on the user studies that demonstrate the superior power of ontology over keyword based search. Next, we are confident that we will be able to develop an intelligent agent that will

dynamically update user profiles. This will provide a level of customization that can have broad application to many areas of content and user interest.

Acknowledgments

This research has been funded in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center with additional support from the Annenberg Center for Communication at the University of Southern California and the California Trade and Commerce Agency.

9. References

[1] S. Adali, K.S. Candan, and S.-S. Chen, K. Erol, and V.S. Subrahmanian. Advanced Video Information System:Data Structures and Query Processing. ACM-Springer Multimedia System, 4:172-186, 1996.

[2] Rune Hjelsvold and Roger Midstraum. Modeling and Querying Video Data. In Proceedings of the 20th International Conference on Very Large Databases (VLDB'94), Santiago, Chile, 1994.

[3] G. Aslan and D. McLeod. Semantic Heterogeneity Resolution in Federated Database by Metadata Implantation and Stepwise Evolution. The VLDB Journal, the International Journal on Very Large Databases, Vol. 18, No. 2, October 1999.

[4] B. Arons. Speech Skimmer: Interactively Skimming Records Speech. Ph.D. Thesis, MIT Media Lab, 1994.

[5] M. A. Bunge. Treatise on Basic Philosophy:Ontology: The Furniture of the World. Reidel, Boston, 1977.

[6] ESPN CLASSIC. http://www.classicsports.com.

[7] CNN Sports. http://www.cnn.com/audioselect.

[8] N. Dimitrova and F. Golshani. Video and Image Content Representation and Retrieval in the Handbook of Multimedia Information Management (editors W. Grosky, R. Jain, and R. Mehrotra), pages 95-138, Prentice Hall, 1998.

[9] T. R. Gruber. Toward Principles for the design of Ontologies used for Knowledge Sharing. In International Workshop on Formal Ontology, March 1993.

[10] Alexander G. Hauptmann. Speech Recognition in the Informedia Digital Video Library: Uses and Limitations. In 7th IEEE International Conference on Tools with AI, Washington, DC, Nov 1995.

[11] Eitetsu Oomoto and Katsumi Tanaka. OVID:Design and Implementation of a Video-Object Database System. IEEE Transactions on Knowledge and Data Engineering, Vol 5, No 4, August 1993.

[12] Alexander G. Hauptmann and M. Witbrock. Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval. Intelligent Multimedia Information Retrieval, Mark T. Maybury, Ed., AAAI Press, pages. 213-239, 1997.

[13] Informix. Informix Universal Server: Informix guide to SQL: Syntax volume 1 & 2 version 9.1, 1997.

[14] David James. The Application of Classical Information Retrieval Techniques to Spoken Documents. Ph.D. Thesis, University of Cambridge, United Kingdom, 1995.

[15] G. J. F. Jones, J. F. Foote, K. Sparck Jone, and S. J. Young. Video Mail Retrieval. In Proc. ICASSP 95, volume I, pages 309-312, Detroit, May 1995.

[16] A. Hampapur and Ramesh Jain. Video Data Management Systems: Metadata and Architecture in Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media (editors Klas, W. and Sheth, A.), chapter 8, pages 245–285. McGraw Hill, 1999.

[17] D.B. Lenat and R.V. Guha. Building Large Knowledge-Based Systems: Representation and Interface in the CYC Project, Addison Wesley, Reading , MA, 1990.

[18] P. Mayes. Agents that Reduce Work and Information Over-load. Communications of ACM, July 1994.

[19] G. Miller. Wordnet: A Lexical Database for English. Communications of CACM, November, 1995.

[20] Large Resources Ontologies (SENSUS) and Lexicons. http://www.isi.edu/naturallanguage/projects/ONTOLOGIES.html

[21] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. Indexing with WordNet Synsets can Improve Text Retreival, Coling-ACL'98 Workshop: Usage of WordNet in Natural Language Processing Systems, pp 38-44, August 98.

[22] L R Rabiner and R.W. Schafer. Digital Processing of Speech Signals., Prentice Hall, 1978.

[23] T.G. A.Smith and N.C. Picever. Parsing Movies in Context. In Proceedings of the 1991 Summer USENIX Conference, Nashville, USA, 1991.

[24] L.D. Wilcox and M.A. Bush. Training and Search Algorithms for an Interactive Wordspotting System. In Proceedings of ICASSP, volume II, pages 97-100, San Francisco, 1992.

[25] Martin Wechsler and Peter Schuble. Metadata for Content-based Retrieval of Speech Recordings in Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media, chapter 8, pages 223–243. McGraw Hill, 1999.

[26] T. Kamba, K. Bharat, and M. Albers. The Krakatoa Chronicle: An Interactive, Personalized Newspaper on the Web. In Proceedings of Fourth International World Wide Web Conference, December, 1995, Boston, MA, USA

[27] CRAYON. http://sun.bucknell.edu/~boulter/crayon

[28] Fishwrap. http://fishwrap.mit.edu/

[29] Ellen Voorhees. Query Expansion Using Lexical-Semantic Relations. 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,1994.

[30] A. F. Smeaton and A. Quigley. Experiments on Using Semantic Distances between Words in Image Caption Retrieval. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1995.