Disambiguation of Annotated Text of Audio 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

1. Background

The development of technology in the field of digital media generates huge amounts of non-textual information, such as audio, video, and images, as well as more familiar textual information. The potential for the exchange and retrieval of information is vast, and at times daunting. In general, users can be easily overwhelmed by the amount of information available via electronic means. The transfer of irrelevant information in the form of documents (e.g. text, audio, and video) retrieved by an information retrieval system and which are of no use to the user wastes network bandwidth and frustrates users. This condition is a result of inaccuracies in the representation of the documents in the database, as well as confusion and imprecision in user queries, since users are frequently unable to express their needs efficiently and accurately. These factors contribute to the loss of information and to the provision of irrelevant information. Therefore, the key problem to be addressed in information selection is the development of a search mechanism which will guarantee the delivery of a minimum of irrelevant information (high precision), as well as insuring that relevant information is not overlooked (high recall).

The traditional solution to the problem of recall and precision in information retrieval employs keyword-based search techniques. Documents are only retrieved if they contain keywords specified by the user. However, many documents contain the desired semantic information, even though they do not contain user specified keywords. This limitation can be addressed through the use of query expansion mechanism. Additional search terms are added to the original query based on the statistical co-occurrence of terms. Recall will be expanded, but at the expense of deteriorating precision. In order to overcome the shortcomings of keyword-based technique in responding to information selection requests we have designed and implemented a concept-based model using ontologies. This model, which employs a domain dependent ontology, is presented in this paper.

1.1 Ontologies

An ontology is a collection of concepts and their interrelationships which can collectively provide an abstract view of an application domain. An ontology defines a set of representational terms that we call *concepts.* Inter-relationships among these concepts describe a target world. Figure 1 shows an example ontology for sports news. This ontology is usually obtained from generic sports terminology and domain experts. This ontology is described by a directed acyclic graph (DAG). Here, each node in the DAG represents a concept. This label name is used to serve as association of concepts with documents. The synonyms list of a concept contains vocabulary (a set of keywords) through which the concept can be matched with user requests. In Figure 1, 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. 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 document’s description (metadata) cannot possess more than one such concept of the disjoint type. For example, when a document’s metadata is the concept “NBA,” it cannot be associated with another disjoint concept, such as “NFL.” Thus, we can say that “NBA,” “CBA,” “college basketball,” and “college football,” are disjoint. Each of these league and its team and player form a boundary what we call *region*.

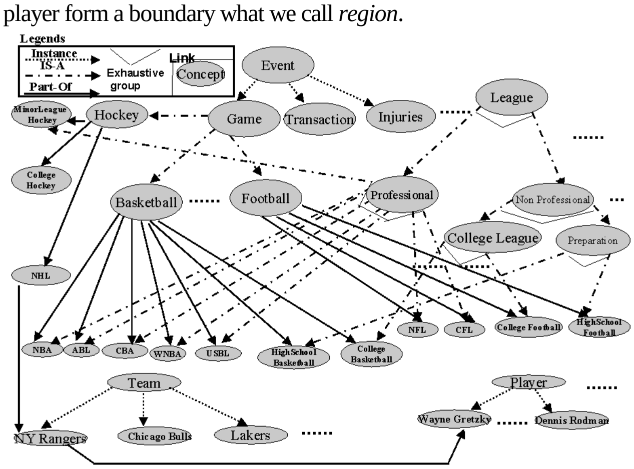

Figure 1 A Small Portion of an Ontology for Sports Domain

There are two distinct questions for ontology-based model: one is the extraction of the semantic concepts from the keywords and the other is the indexing. With regard to the first problem, the key issue is to identify appropriate concepts that describe and identify documents on the one hand, and on the other, the language employed in user requests. In this it is important to make sure that irrelevant concepts will not be associated and matched, and that relevant concepts will not be discarded. In other words, it is important to insure that high precision and high recall will be preserved during concept selection for documents or user requests. In this paper, we propose an automatic mechanism for the selection of these concepts from documents by addressing the first problem. Furthermore, this concept selection mechanism includes a novel, scalable disambiguation algorithm using domain specific ontology. This algorithm will prune irrelevant concepts while allowing relevant concepts to become associated with documents and participate in query generation.

In order to demonstrate the effectiveness of our disambiguation model we have explored and provided a specific solution to the problem of retrieving audio information. The effective selection/retrieval of audio information entails several tasks, such as metadata generation (description of audio), and the consequent selection of audio information in response to a query. Relevant to our purpose, ontologies can be fruitfully employed to facilitate metadata generation. For metadata generation, we need to do content extraction by relying on speech recognition technology which converts speech to text. After generating transcripts we can deploy our ontology-based disambiguation algorithm to identify appropriate concepts that describe and identify audio documents and serve as metadata for the documents.

2. Disambiguation Algorithm

For each audio document or object we need to find the most appropriate concept(s) from associated keywords that appear in a given audio object. For this, concepts from ontologies will be selected based on matching terms taken from their lists of synonyms with those based on specified keywords. It is possible that a particular keyword may be associated with more than one concept in the ontology. In other words, association between keyword and concept is one:many, rather than one:one. Therefore, the disambiguation of concepts is required. The basic notion of disambiguation is that a set of keywords occurring together determine a context for one another, according to which the appropriate senses of the word (its appropriate concept) can be determined. Note, for example, that base, bat, glove may have several interpretations as individual terms, but when taken together, the intent is obviously a reference to baseball. Our disambiguation algorithm is based on two principles: co-occurrence and semantic closeness. This algorithm first strives to disambiguate across several regions using first principle, and then disambiguates within a particular region using the second. For automatic disambiguation within an ontology a set of regions representing different concepts can be defined. The concepts, as they appear in a given region, will be mutually disjoint from the concepts of other regions. This becomes the basis for determining a group of appropriate concepts for a given keyword or collection of keywords. In short, after keywords are matched to the concepts of a given ontology, the region within the ontology in which the greatest number of selected concepts occurs is determined. This region, the one containing the largest number of selected concepts, will at the time be used to associate with documents. The selected concepts of other, different, regions will be pruned automatically.

We will need to have a way of dealing with the possibility that even within a region selected for annotation a given keyword will match more than one concept. In other words, within a given region multiple ambiguous concepts will have been selected for a particular keyword, necessitating further disambiguation. In order to further prune irrelevant concepts we will need to determine the correlation between concepts selected in a given region. For this, we use the second principle; semantic distance (in the ontology). When concepts are correlated, concepts closely associated will be given greater weight. This association will be based on minimal distance in the ontology and the matching scores of concepts based on the number of keywords they match. Thus, selected concepts which correlate with each other will have a higher score, and a greater probability of being retained than non-correlated concepts. If scores of particular ambiguous concepts fall below a certain *threshold-score*, which will be a minimum score chosen for selected concepts for that particular object, these concepts will be pruned. For example, the annotated text for a particular audio object might be: "*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 words in Italics are the keywords which are associated with the concepts of our ontology. The keywords "Lakers," and "Nets" are associated with the concepts "Los Angeles Lakers" and "New Jersey Nets" respectively. The keyword "Bryant" is associated with the concepts, "Reeves Bryant," "Bryant Mark," and "Bryant Kobe."

It is important to note that all the concepts selected above are found in the region "NBA." However, the keyword "Eastern" is associated with the concepts "Eastern Washington," and "Eastern Michigan" which are not associated with "NBA" but with the region "College Basketball." If we now choose only the concepts which appear in the region NBA, which is the region in which the greatest number of concepts occur, the concepts "Eastern Washington" and "Eastern Michigan" will be eliminated, since they are not found in that region. Thus, we keep from among the concepts selected those which appear in the region NBA in which the greatest number of concepts occur, and prune other selected concepts. In the selected region, in this case NBA, a keyword such as "Bryant" may be associated with more than one selected concept. This necessitates further disambiguation. We will want to know what other concept qualifies the concepts selected by keyword "Bryant" through correlation. As noted above in the case of the keyword "Bryant" the concepts "Bryant Kobe," "Bryant Mark," and "Reeves Bryant" are all selected. Among these ambiguous concepts, however, only "Bryant Kobe" is correlated with another selected concept, in this case "Los Angeles Lakers." Therefore, "Bryant Kobe" is kept, and the concepts "Bryant Mark," and "Reeves Bryant" are thrown away. Finally, the selected concepts for this audio object are "New Jersey Nets," "Los Angeles Lakers," and "Bryant Kobe."

At present, an experimental prototype for the implementation of the model has been developed and implemented. As of today, our working ontology has around 7,000 concepts for the sports news domain, with 2,481 audio clips/objects of metadata in the database. For sample audio content we use CNN broadcast sports and Fox Sports audio, along with closed captions. Using our disambiguation algorithm, these associated closed captions are connected with the ontology. The performance of our disambiguation algorithm has been studied by considering what percentage of the audio objects selected are associated with relevant concepts. We have observed that through the use of our disambiguation algorithm 90.5% of the objects extrapolated from closed captions successfully associate with concepts of ontologies, while only 9.5% of the objects fail to associate. Among these 90.5%, up to 76.9% of the objects are associated with relevant concepts (pure); in other cases, objects are associated with relevant and irrelevant concepts (mixed).