

Data Semantics for Improving Retrieval Performance of Digital News Video Systems¹

G. Ahanger and T.D.C. Little

Department of Electrical and Computer Engineering
Boston University, Boston, Massachusetts 02215, USA
(617) 353-9877
{gulrukh,tdcl}@bu.edu

MCL Technical Report No. 07-10-1999

Abstract—We propose a novel four-step hybrid approach for retrieval and composition of video newscasts based on information contained in different metadata sets. In the first step, we use conventional retrieval techniques to isolate video segments from the data universe using segment metadata. In the second step, retrieved segments are clustered into potential news items using a dynamic technique sensitive to the information contained in the segments. In the third step, we apply a transitive search technique to increase the recall of the retrieval system. In the final step, we increase recall performance by identifying segments possessing creation-time relationships.

A quantitative analysis of the performance of the process on a newscast composition shows an increase in recall by 59% over the conventional keyword-based search technique used in the first step.

Keywords: News video composition, retrieval, content metadata, structural metadata, unstructured metadata, keyword vector, recall, precision.

¹*IEEE Trans on Knowledge and Data Engineering*, Vol. 13, No. 3, May/June 2001, pp. 353-360. This work is supported in part by the National Science Foundation under Grant No. IRI-9502702. Part of this work was presented at the 8th IFIP 2.6 Working Conference on Database Semantics, Rotorua, New Zealand, January 1999.



Figure 1: Scenes from an Example News Item

1 Introduction

A challenging problem in video-based applications is achieving rapid search and retrieval of content from a large corpus. Because of the computational cost of real-time image-based analysis for searching such large data sets we pursue techniques based on off-line or semi-automated classification, indexing, and cataloging. Therein lies the need for “bridge” techniques that have rich semantics for representing motion-image-based concepts and content, yet are supported by fast and efficient algorithms for real-time search and retrieval. At this intersection we have been investigating techniques for video concept representation and manipulation. In particular we have sought the goal of automatic composition of news stories, or newscasts based on an archive of digital video with supporting metadata.

The general process for automatic composition of digital news video (or other video type) towards a theme is based on selecting desired video data within some domain (e.g., sports), filtering redundant data, clustering similar data in sub-themes, and composing the retrieved data into a logical, chronological, and thematically-correct order [2]. All of these tasks are possible if sufficient information is available about the content of the video data. Therefore, information (metadata) acquisition and techniques to match, filter, and compose video data are critical to the performance of a video composition system. The quality (precision) of data retrieved depends on the type of metadata and the matching technique used.

However, news audio and video (and associated closed-captioning) do not necessarily

Table 1: Transcripts of Several Segments

<i>Introduction</i>	<i>Field Scene</i>	<i>Interview</i>
A ONE-YEAR-OLD BABY BOY IS SAFE WITH HIS MOTHER THIS MORNING, THE DAY AFTER HIS OWN FATHER USED HIM AS A HOSTAGE. POLICE SAY IT WAS A DESPERATE ATTEMPT TO MAKE IT ACROSS THE MEXICAN BORDER TO AVOID ARREST. CNN'S ANNE MCDERMOTT HAS THE DRAMATIC STORY.	A MAN EMERGED FROM HIS CAR AT THE U.S. MEXICAN BORDER, CARRYING HIS LITTLE SON, AND A KNIFE. WITNESSES WITNESSES SAY HE HELD THE KNIFE TO HIS SON, LATER, TO HIMSELF. AND IT ALL PLAYED OUT LIVE TV. OFFICIALS AND POLICE FROM BOTH SIDES OF THE BORDER...	DARYN: JUST IN THE RIGHT PLACE AT RIGHT TIME ESPECIALLY FOR THIS LITTLE BABY. CAN YOU TELL US WHAT YOU WERE SAYING TO THE MAN POLICE IDENTIFIED AS EDDIE PRICE AND WHAT HE WAS SAYING ON BACK TO YOU? I JUST ASSURED HIM THAT THE BABY WOULD BE OKAY...

possess correlated concepts (Fig. 1). For example, it is common in broadcast news items that once an event is introduced, in subsequent segments the critical keywords are alluded to and not specifically mentioned (e.g., Table 1, the name “Eddie Price” is mentioned only in the third scene). Segments can share other keywords and can possess transitive relationships. If a search is performed on a person’s name, then all related segments are not necessarily retrieved. Similarly, related video segments can have different visuals. It is not prudent to rely on a single source of information about the segments in retrieval and composition (e.g., transcripts or content descriptions). The information tends to vary among the segments related to a news item. Therefore, we require new techniques to retrieve all the related segments in a video composition system (i.e., improve the recall [22]). In this paper, we use the information contained in audio, transcripts, and closed-caption data interchangeably. All three formats represent spoken text within video clips.

We also propose a transitive video composition and retrieval approach that improves recall. That is, once a query is matched against unstructured metadata (e.g., closed-caption and speech-to-text), the components retrieved are again used as queries to retrieve additional video segments with information belonging to the same news item. The recall performance can be further enhanced if the union of different metadata sets is used to retrieve all segments of a news item (Fig. 2). However, the union operation does not always guarantee full recall as a response to a query. This is because no segment belonging to a particular instance of a news item may be present among the segments acquired after the transitive search (data acquired from different sources or over a period of time containing data about the same news event).

Our work is an outcome of observations of generative semantics in the different forms of information associated with news video data. The information can be in the visuals or in the audio associated with the video. We also study the common bond among the

Table 2: Content Metadata

Entity	Tangible object that are part of a video stream. The entities can be further sub-classified, (e.g., persons, and vehicles).
Location	Place shown in video. (e.g., place, city, and country).
Event	Center or focus of a news item.
Category	Classification of news items.

Table 3: Structural Metadata

1. Headline	Synopsis of the news event.	
2. Introduction	Anchor introduces the story.	
3. Body	Describes the existing situation.	
	a. Speech	Formal presentation of views without any interaction from a reporter.
	b. Comment	Informal interview of people at the scene in the presence of wild sound.
	c. Wild Scene	Current scenes from the location.
	d. Interview	One or more people answering formal structured questions.
	e. Enactment	Accurate scenes of situations that are already past.
4. Enclose	Contains the current closing lines.	

segments belonging to a single news item. The composition should possess a smooth flow of information with no redundancy.

Annotated metadata are the information extracted from video data. In our previous work [4, 18] we have classified annotated metadata that are required for a newscast composition as content metadata and structural metadata. The content metadata organize unstructured information within video data (i.e., objects and interpretations within video data or across structural elements). Some of the information extracted from news video data is shown in Table 2. Information such as the objects present in visuals, the category of a news item, and the main concept (focus or center [8]) depicted by the new item are stored as metadata. The structural metadata organize linear video data for a news item into a hierarchy [3] of structural objects as shown in Table 3. Therefore, each structural object is represented by content metadata.

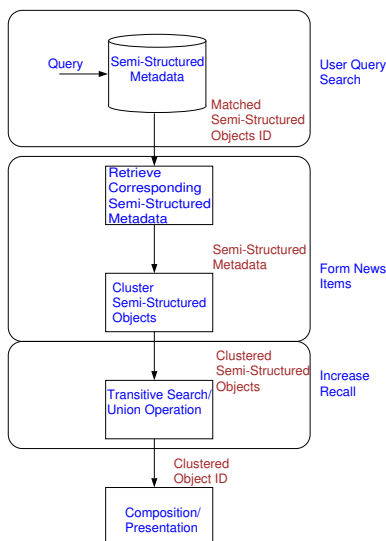


Figure 2: Process Diagram for Newscast Video Composition

The development of the proposed hybrid video data retrieval technique is based on the availability of segment metadata. We have explored the use of these data for the following reasons:

- By utilizing both annotated metadata and closed-caption metadata, precision of the composition system increases. For example, keywords of “Reno, Clinton, fund, raising,” if matched against closed-caption metadata, can retrieve information about a place called “Reno” (Nevada). Therefore, annotated metadata can be used to specify that only a person called “Reno” (Janet Reno) should be matched. The results from annotated and closed-captioned searching can be intersected for better precision.
- Recall of a keyword-based search improves if more keywords associated with an event are used. Transcripts provide enriched but unstructured metadata, and can also be used to improve recall. Utilizing transcripts increase the number of keywords in a query; therefore, in some cases precision of the results will be compromised (irrelevant data are retrieved). The transitive search technique is based on this principle (Section 4).
- If the relationships among segments of a news event are stored, recall of a system can be increased. For example, if news about “Clinton” is retrieved, then related segment types can be retrieved even if the word “Clinton” is not in them.

As a result of the above observations, we propose a hybrid approach that is based on

the union of metadata sets and keyword vector-based clustering as illustrated in Fig. 2. The precision of vector-based clustering improves by using multiple indexing schemes and multiple sets of metadata (annotated and unstructured). Unstructured data describe loosely organized data such as free-form text of the video transcripts.

The organization of the remainder of this paper is as follows: In Section 2 we describe existing techniques for video data retrieval. In Section 3 we discuss metadata required for query processing, classification of annotated metadata, and the proposed query processing technique. In Section 4 we present an analysis of the proposed approach. Section 5 presents our observations of the process, and Section 7 concludes the paper.

2 Related Work in Video Information Retrieval

A variety of approaches have been proposed for the retrieval of video data. They can be divided into annotation-metadata-based, transcript-metadata-based, and hybrid-metadata-based techniques. Each is described below.

For annotation-based techniques, manual or automatic methods are used for extraction of information contained in video data. Image processing is commonly used for information extraction in the automatic techniques. Techniques include automatic partitioning of video based on information within video data [5], extraction of camera and object motion [6, 24], and object, face, texture, visual text identification [7, 12, 16, 19, 20, 21, 23]. The metadata describing large digital video libraries can also be extracted off-line and stored in a database for fast query processing and retrieval [7].

Transcripts associated with video data can provide an additional source of metadata associated with video segments. Brown et al. [9] use transcript-metadata to deliver pre-composed news data. Wachman [25] correlates transcripts with the scripts of situation comedies. The Informedia project [26] uses a hybrid-metadata approach to extract video segments for browsing using both the visual and transcript metadata.

In the above works, keyword searching is either used to retrieve a pre-assembled news item or the segments associated with the query keywords. The objective in our work is to search for segments that belong to the various instances of the same event and to cover various time periods (e.g., retrieve information about Albright’s trip to the Middle East). Therefore, we seek to maximize the availability of information to support the creation of a cohesive video piece. For this purpose we require, in addition to the the segments matching

a query, any segments that are related via a transitive or structural relationship. In this manner, segments belonging to various instances of a news event can be merged to create a new composition. Our technique uses a four-step approach applied to both annotation-based (content and structural) and transcript-based (unstructured) metadata. We use a transitive search on transcripts and the union operation on structural metadata to retrieve related video segments.

3 The Proposed Four-Step Hybrid Technique

The four-step hybrid retrieval technique is based on establishing transitive relationships among segment transcripts and the use of annotated metadata. After introducing our terminology (symbols used throughout the paper are summarized in Table 4), we describe the different types of metadata and how they are used to support the four-step process.

3.1 Preliminaries

Metadata described in this paper include unstructured metadata such as free-form text and annotation metadata. The former are used for transitive search. The latter are comprised of content metadata and structural metadata.

Unstructured Metadata and Transitivity Transcripts originating from closed-caption data (audio transcripts), when available, are associated with video segments when the segments enter the content universe S . These transcripts comprise the unstructured metadata for each segment.

Unstructured metadata are used for indexing and forming keyword vectors for each semi-structured metadata segment. Indexing is the process of assigning appropriate terms to a component (document) for its representation. Transitivity on the unstructured data is defined below.

Let \mathcal{R}_f define a binary relationship f on the universal set of video segments S (i.e., $(s_a, s_b) \in R_f \iff s_a$ is *similar* to s_b). If similarity distance, defined as $d(s_a, s_b)$ for segments s_a and s_b , is greater than an established value then the two segments are considered to be similar. The transitive search satisfies the following property (for all $s_a \in S, s_b \in S, s_c \in S$):

Table 4: Symbols Used to Define the Retrieval Technique

Symbols	Descriptions
s	A video segment
S	Universe of video segments
N	Size of the universe S
R_f	A binary relationship on S for transitive search
R_u	A binary relationship on S for related segment search
tf_i	Frequency of a concept (term) i in unstructured metadata
N_i	Number of unstructured metadata components with term i
w_{1_i}	Intermediate weight assigned to a concept i for query match
w_{2_i}	Final weight assigned to a concept i for query match
w_{3_i}	Final weight assigned to a concept i for transitive search
q	A query
S_q	A set of segments returned as a result of a query
$d(s_a, s_b)$	The similarity distance between two sets of keywords
QS	A subset of S_q
T_c	Cluster cut-off threshold
CL_i	A cluster
$q(s)$	A query comprised of unstructured metadata component
s_t	A segment retrieved as a result of a query $q(s)$
$S_{q(s)}$	Set of segments s_t retrieved as a result of a query $q(s)$
TCL_i	An extended cluster CL_i resulting from a transitive search
S_a	A candidate set resulting from cluster TCL_i

$$(s_a, s_b) \in \mathcal{R}_f \wedge (s_b, s_c) \in \mathcal{R}_f \Rightarrow (s_a, s_c) \in \mathcal{R}_f$$

Therefore, for a transitive search we first match a query with unstructured metadata in the universe S . The results are applied as a query to retrieve additional unstructured metadata (transcripts) and associated segments, increasing the the recall of the process.

Annotated Metadata Annotated metadata consist of content and structural metadata as described in Section 1. Structural metadata exist if segments are annotated as such when they enter the segment universe, S , either as video shot at a single event (e.g., a sporting event) or as decomposed segments originating from pre-assembled news items (as is the case for our dataset). We call such segments *siblings* if they posses either of these relationships.

A shortcoming of the aforementioned transitive search is that it may not retrieve all

segments related via siblings. This can be achieved by the following.

Let \mathcal{R}_u define a binary relationship u on the universal set S (i.e., $(s_a, s_b) \in \mathcal{R}_u \iff s_a$ and s_b are part of the same news event). The final step expands the set of segments as a union operation as follows:

$$S_a \leftarrow S_a \cup \{s_b \mid \exists s_a \in S_a : (s_a, s_b) \in \mathcal{R}_u\},$$

where, S_a represents the candidate set of segments used as a pool to generate the final video piece (or composition set) [2].

The hierarchical structure of related segments is stored as structural metadata that are utilized in the proposed hybrid retrieval technique (Table 3).

3.2 Segment Keyword Analysis and Weighting

We use text indexing and retrieval techniques proposed by Salton [22] and implemented in SMART [10] for indexing the unstructured metadata. To improve recall and precision we use two sets of indices, each using different keyword/term weighting. In the remainder of the paper we use s interchangeably to represent a video segment or its associated unstructured metadata. The similarity distance of a segment with a query or a segment is measured by the associated unstructured metadata.

The selection process is comprised of an initial segment weighting followed by a clustering step.

Initial Segment Weighting Initially, a vector comprised of keywords and their frequency (term frequency tf) is constructed using the unstructured metadata of each segment without stemming and without common words. The frequency of a term or keyword indicates the importance of that term in the segment. We then normalize the tf in each vector with segment (document) frequency in which the term appears by using Eq. 1.

$$w_{1_i} = tf_i \times \log \left(\frac{N}{N_i} \right)^2, \quad (1)$$

where N is the number of segments in the collection, and N_i represents the number of

segments to which term i is assigned. The above normalization technique assigns a relatively higher weight w_{1_i} to a term that is present in smaller number of segments with respect to the complete unstructured metadata. Finally, w_{1_i} is again normalized by the length of the vector (Eq. 2). Therefore, the influence of segments with longer vectors or more keywords is limited.

$$w_{2_i} = \frac{w_{1_i}}{\sqrt{\sum_{j=0}^n (w_{1_j})^2}} \quad (2)$$

Clustering and Transitive Weighting In this step we use word stemming and stop words to increase search sensitivity to variants of the same keyword. In segments belonging to a news item, the same word can be used in multiple forms. Therefore, by stemming a word we achieve a better match between segments belonging to the same news item. For the transitive search and clustering, we use the complete unstructured metadata of a segment as a query, resulting in a large keyword vector because we want only the keywords that have a high frequency to influence the matching process. Therefore, we use a lesser degree of normalization (Eq. 3) as compared to the initial segment weighting.

$$w_{3_i} = tf_i \times \log \left(\frac{N}{N_i} \right) \quad (3)$$

Table 6 shows a comparison of the weighting schemes for the same unstructured metadata. The two concepts “Iraq” and “Iraqi” in the second scheme are treated as the same and hence the concept “Iraq” gets a higher relative weight.

For the purpose of a query match we use the cosine similarity metric (Eq. 4) proposed by Salton. The metric measures the cosine or the measure of angle between two unstructured metadata segment vectors. The product of the length of the two segment vectors divides the numerator in the cosine metric. The longer length vectors produce smaller cosine similarities. n represents the number of terms or concepts in the universe in Eq. 4.

$$\text{cosine}(\vec{A}, \vec{B}) = \frac{\sum_{k=1}^n (a_k \times b_k)}{\sqrt{\sum_{k=1}^n (a_k)^2 \times \sum_{k=1}^n (b_k)^2}} \quad (4)$$

The proposed query processing technique is a bottom-up approach in which the search begins using the unstructured metadata. We describe the details next.

3.3 The Selection Mechanism

The four-step selection mechanism is illustrated Fig. 2. A query enters the system as a string of keywords. These keywords are matched against the indices created from the unstructured metadata. The steps of this process are query matching, clustering the results, retrieval based on the transitive search, and sibling identification. These are described below.

Query Matching This stage involves matching of a user-specified keyword vector with the available unstructured metadata. In this stage we use indices that are obtained as a result of the initial segment weighting discussed in the previous section. Because the match is ranked-based, the segments are retrieved in the order of reduced similarity. Therefore, we need to establish a cut-off threshold below which we consider all the segments to be irrelevant to the query. Unfortunately, it is difficult to establish an optimal and static query cut-off threshold for all types of queries as the similarity values obtained for each query are different. For example, if we are presented with a query with keywords belonging to multiple news items, then the similarity value with individual object in the corpus will be small. If the query has all keywords relevant to single news item then the similarity value will be high. Because of this observation, we establish a dynamic query cut-off threshold ($D \times \max\{d(s, q)\}$) and we set it as a percentage D of the highest match value $\max\{d(s, q)\}$ retrieved in set S_q . The resulting set is defined as:

$$QS \leftarrow \{s \in S_q \mid d(s, q) \geq (D \times \max\{d(s, q)\})\},$$

where s is the segment retrieved and $d(s, q)$ is the function that measures the similarity distance of segment s returned as a result of a query q .

Results Clustering In this stage, we cluster the retrieved segments with each group containing yet more closely related segments (segments belonging to the same event). We use the indices acquired as a result of the transitive scheme (Fig. 3). During the clustering process, if the similarity ($d(s_a, s_b)$) of the two segments is within a cluster cut-off threshold T_c , then the two segments are considered similar and have a high probability of belonging to the same news event. Likewise, we match all segments and group the segments that have similarity value within the threshold, resulting in a set

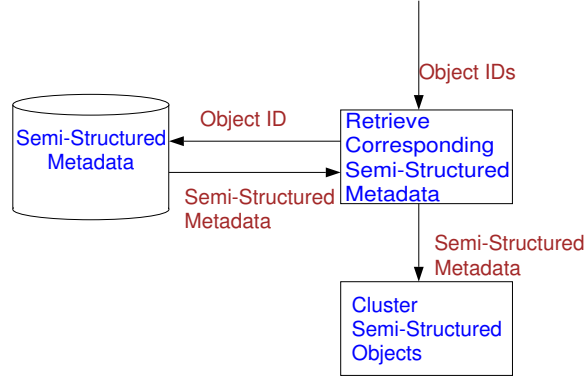


Figure 3: Process Diagram of the Clustering Process

$$\{CL_1, CL_2, CL_3, \dots, CL_k\},$$

where CL_i are clusters (sets) each consisting of segments belonging to a single potential news item. An algorithm for forming the clusters is described below.

For forming disjoint clusters we use a graph-theoretic method [15, 17] that uses minimal spanning tree (MST). The longest edges in the tree are removed producing clusters. We use a threshold T_c (the edges with length beyond and equal to which are removed) that gives the best clustering performance on the experimental data set. However, if an optimum threshold is to be used, then the cluster separation measure proposed by Davies and Bouldin [14] can be used. For creating the MST we use Prim's algorithm [13] and a depth-first search algorithm to find long edges in the tree. This type of search is used due to ease with which the clusters are created. The clusters are formed as follows:

1. If there are k segments in the set QS then first create the $k \times k$ similarity matrix $F = [f_{ij}]$, where

$$f_{ij} = \begin{cases} \frac{1}{d(s_i, s_j)} & \text{if } i \neq j \wedge d(s_i, s_j) > 0 \\ 0 & \text{if } i \neq j \wedge d(s_i, s_j) = 0 \\ 0 & \text{if } i = j \end{cases} \quad i, j = 1, \dots, k$$

2. Use Prim's algorithm for forming MST. The input to the algorithm is the matrix F and the output is the tree.
3. Use a depth-first traversal through the tree to remove edges greater than the threshold T_c . This results in separate clusters CL_i of connected nodes.

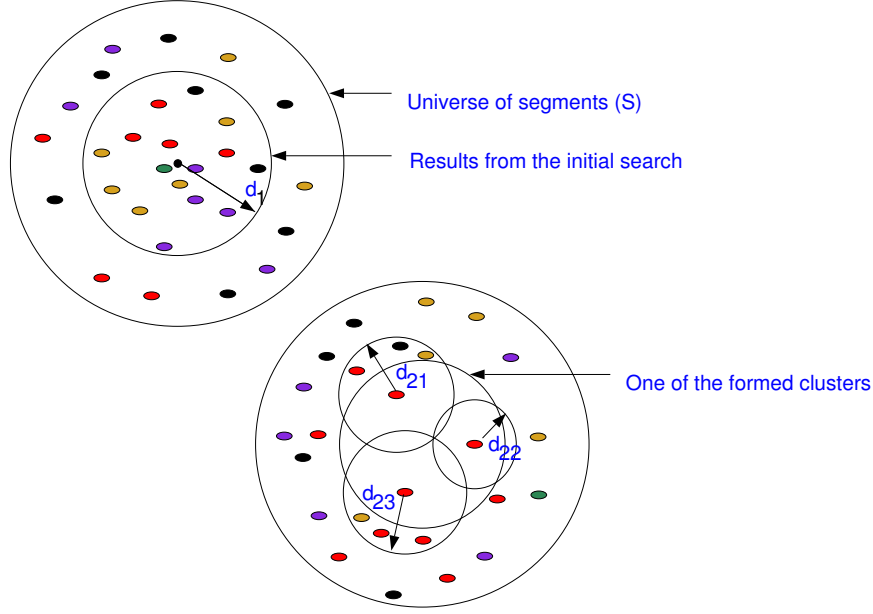


Figure 4: Similarity Measure based on the Transitive Search

Transitive Retrieval We use a transitive search (Fig. 4). The transitive search increases the number of segments that can be considered similar. During query matching, the search is constrained to the similarity distance (d_1), thereby only retrieving segments within this distance. During the transitive search we increase the similarity distance of the original query by increasing the keywords in the query so that segments within a larger distance can be considered similar. In the transitive search we use unstructured metadata of each object in every cluster as a query, $q(s)$, and retrieve similar segments. Again, item cut-off threshold is used as a cut-off point for retrieved results and the retained segments are included in the respective cluster.

The transitive cut-off threshold ($T \times \max\{d(s_t, q(s))\}$) is set as the percentage (T) of the highest similarity value retrieved $\max\{d(s_t, q(s))\}$. For example, the distances d_{21} , d_{22} , and d_{23} (Fig. 4) fall within the transitive cut-off thresholds of respective segments.

Consider a cluster $CL_i = \{s_1, s_2, s_3, \dots, s_N\}$ formed in the results clustering step. The extended cluster resulting from the transitive search can be defined as:

$$TCL_i \leftarrow \bigcup_{\forall s \in CL_i} \{s_t \in S_{q(s)} \mid d(s_t, q(s)) \geq (T \times \max\{d(s_t, q(s))\})\},$$

where, s_t is a segment returned as a result of a transitive search of a segment $s \in CL_i$,

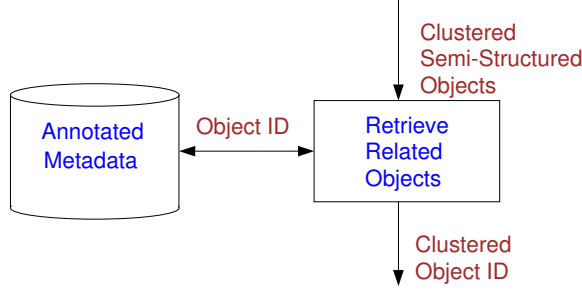


Figure 5: Process Diagram for Retrieving Related Segments

$d(s_t, q(s))$ is the function that measures the similarity value of a segment s_t to query $q(s)$.

Sibling Identification To further improve recall we use the structural metadata associated with each news item to retrieve all other related objects (Fig. 5). Structural information about each segment in a cluster is annotated; therefore, we have the information about all the other segments that are structurally related to a particular segment. We take the set of segments that are structurally related to a segment in a cluster and perform a union operation with the cluster. Suppose $TC_i = \{s_1, s_2, s_3, \dots, s_n\}$ is one of the cluster resulting from the third step. The final set can be defined as:

$$SC_i = \bigcup_{s \in TC_i} R(s)$$

Here $R(s)$ is a set of segments related to the segments s . Likewise, the union operation can be performed on the remaining clusters.

By using this four-step hybrid approach we are able to increase the recall of the system. Next we discuss the performance of the retrieval, clustering, and proposed transitive search process.

4 Analysis of the Proposed Hybrid Technique

We evaluated the performance of our technique based on 10 hours of news video data and their corresponding closed-caption data acquired from network sources. Our results and analysis of the application of our techniques on this data set are described below.

Because the objective of our technique is to yield a candidate set of video segments suitable for composition, we focus on the inclusion-exclusion metrics of recall and precision for evaluating performance. However, subsequent rank-based refinement on the candidate set yields a composition set that can be ordered for a final video piece [2].

The data set contains 335 distinct news items obtained from CNN, CBS, and NBC. The news items comprise a universe of 1,731 segments, out of which 1,486 segments are relevant to the queries executed. The most common stories are about bombing of an Alabama clinic, Oprah Winfrey’s trial, the Italian gondola accident, the UN and Iraq standoff, the Pope’s visit to Cuba, and the Clinton controversy. The set of keywords used in various combinations in query formulation is as follows:

race relation cars solar planets falcon reno fund raising
 oil boston latin school janet reno kentucky paducah rampage
 santiago pope cuba shooting caffeine sid digital genocide
 compaq guatemala students chinese adopted girls
 isreal netanyahu arafat fda irradiation minnesota tobacco trial
 oprah beef charged industry fire east beach varadero
 gay sailor super bowl john elway alabama clinic italy
 gondola karla faye tucker dead advertisers excavation lebanon
 louise woodward ted kaczynski competency birmingham islam blaze
 vegetarian police maryland commercials teachers mir undocking soyuz
 marine ski resort cable federal north carolina peta hamburger
 plants strike coca cola medicare biological weapons warheads

The number of keywords influences the initial retrieval process for each news item used in a query. If more keywords pertain to one news item than the other news items, the system will tend to give higher similarity values to the news items with more keywords. If the query cut-off threshold is high (e.g., 50%), then the news items with weaker similarity matches will not cross the query cut-off threshold (the highest match has a very high value). Therefore, if more than one distinct news item is desired, a query should be composed with equal number of keywords for each distinct news item. All of the distinctly retrieved news items will have approximately the same similarity value to the query and will cross the query cut-off threshold.

For the initial experiment we set the query cut-off threshold to 40% of the highest value retrieved as a result of a query, or $0.4 \times \max(S_q)$. The transitive cut-off threshold was set to 25% of the highest value retrieved as a result of unstructured metadata query, or $0.25 \times \max(S_q(s))$. The results of 50 queries issued to the universe are shown in Figs. 6 and 7. Here we assume that all the segments matched the query (we consider every retrieved segment a positive match because the segments contain some or all keywords of the query).

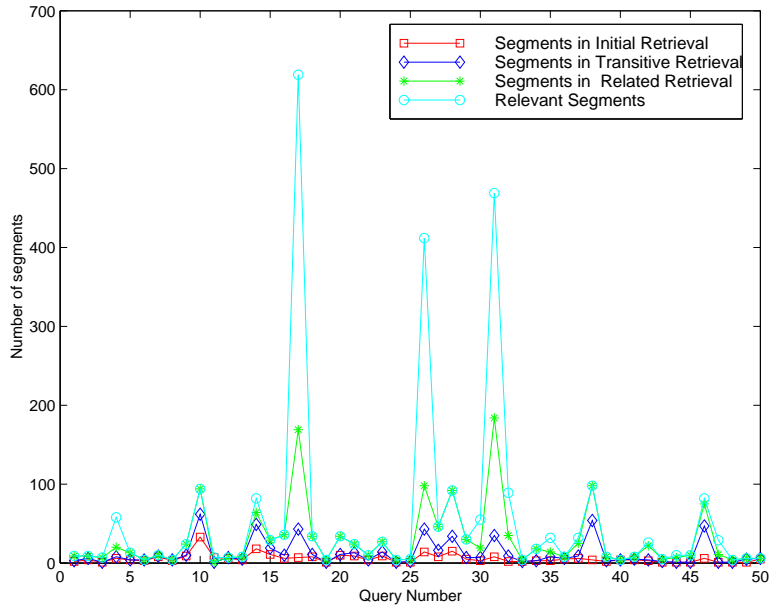


Figure 6: Summary of Performance of Different Retrieval Techniques

Not all of the keywords are common among the unstructured metadata of related segments, nor are they always all present in the keywords of a query. Therefore, to enhance the query we use a transitive search with a complete set of unstructured metadata. The probability of a match among related segments increases with the additional keywords; however, this can reduce precision.

As the result of the transitive search the recall of the system is increased to 42% from 29% (another iteration of transitive search can increase it further) as shown in Fig. 8. The range of increase in recall among the 50 queries is between 0% and 51%. Recall increased by 46% as a result of the sibling identification.

The precision of the results due to the transitive search is reduced to 90% from 100% as shown in Fig. 9. The precision is further reduced to 85% in the sibling identification step. The reduction in precision is due to irrelevant segments retrieved during the transitive search.

A cause of such low recall of the initial retrieval and subsequent transitive search is the quality of the unstructured metadata. Often this quality is low due to incomplete or missing sentences and misspelled words (due to real-time human transcription).

Using the structural hierarchy (Section 3.1) we capture the relationships among the segments belonging to a news item. Therefore, if this information is exploited we can get

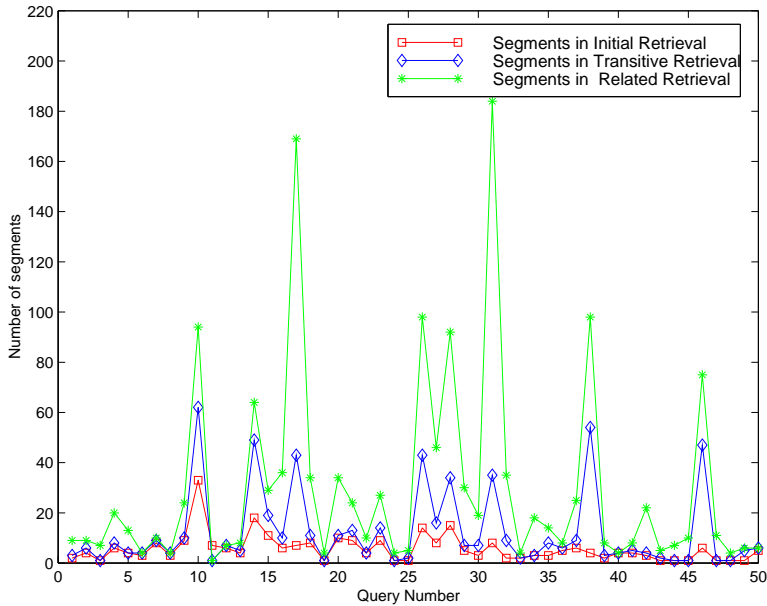


Figure 7: Improvement in Performance of Transitive and Sibling Retrieval Techniques

an increase in recall without a reduction in precision (as all segments belong to the same news item). In the last step of the query processing we use structural metadata to retrieve these additional segments. As observed from the above results, the recall is then increased to 96%. The remaining data are not identified due to a failure of the prior transitive search.

The results demonstrate that the combination of different retrieval techniques using different sources of metadata can achieve better recall in a news video composition system as compared to a the use of a single metadata set.

5 Observations

To emulate news items which encompass multiple foci (i.e., concepts from each are associated with many segments), it becomes difficult to balance the clustering of segments for these foci with our techniques. For example, the query “State of the Union Address” applied to our data set will yield foci for the address and the intern controversy. However, there are many more segments present in the data set for the intern controversy.

The query precision can also be increased by forming the intersection of the keywords from the content and unstructured metadata sets. For example, consider the scenario for composing a news item about Clinton speaking in the White House about the stalemate in

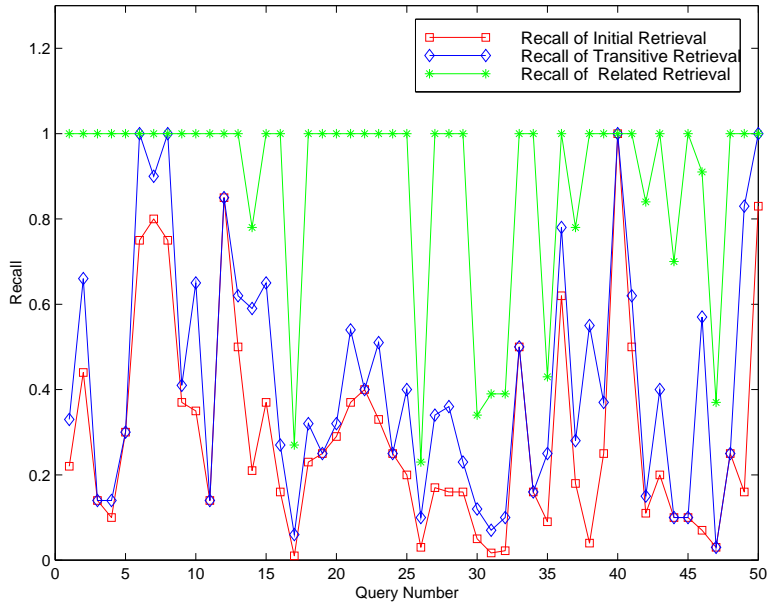


Figure 8: Recall of Different Retrieval Techniques

the Middle East. From the content metadata, we might be able to retrieve segments of type **Speech** for this purpose. However, many of the returned segments will not be associated with the topic. In this case an intersection of the query results of the salient keywords applied to the unstructured metadata will give us the desired refinement (Fig. 10).

If a query retrieves a set of new items based on a date or period then access can be achieved directly from the content metadata. For the process of composition, the broader set of metadata needs to be used.

6 Implementation

To evaluate the effectiveness of the proposed hybrid retrieval technique, we implemented a news digital video production system (DVPS). A Web-based version of this prototype, called *Canvass*, is currently functional [1]. The architecture of the system implementation is shown in Fig. 11. Various technologies have been integrated to develop the DVPS and these are be discussed as part of the system architecture.

In the four step hybrid approach for retrieval of video data, we have utilized the semantics within unstructured data to aid in retrieval. However, visual data is perceived by different users differently and unstructured metadata does not capture a user context adequately.

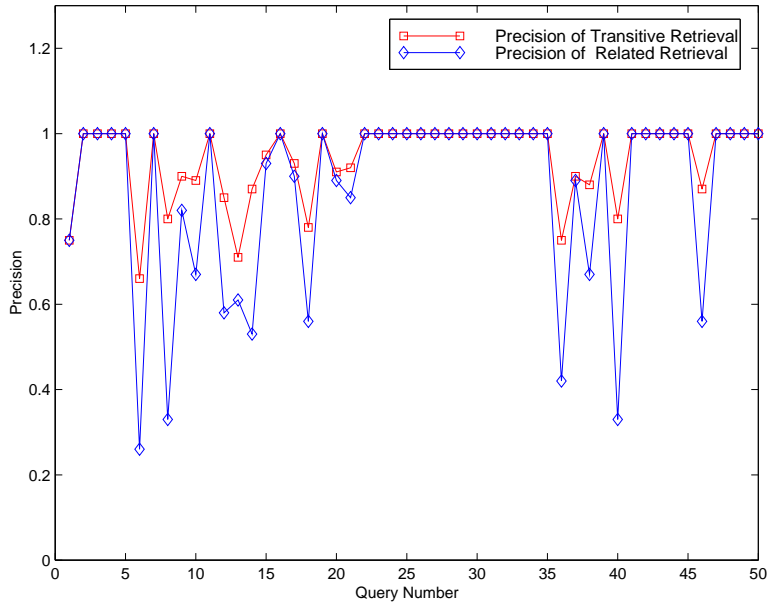


Figure 9: Precision of the Transitive and Sibling Retrieval Techniques

Therefore, in the DVPS we utilize annotated metadata. For example, a user query can be first matched with annotated metadata and the unstructured metadata associated with the segments thus retrieved are used to improve the recall. First, we use the transitive search on the segments (unstructured metadata) retrieved and then use the sibling relationship to further improve the recall.

The architecture of the news DVPS is divided into an annotation and indexing module, a retrieval and composition module, and a video delivery module. The annotation and indexing module is used to annotate and index video data and transcripts, respectively, and to populate a database with the annotated metadata. In the retrieval module we process queries using both annotated and transcript metadata and compose the resulting data. The proposed hybrid retrieval technique and composition techniques [2] are implemented in this module. The video delivery module is used to schedule the playout of actual video segments within a composition.

To use the system, analog video data are first converted into a digital format (currently MPEG 1 and Real formats). The digital video are annotated using Vane. The output of Vane, raw content metadata, are stored in a SGML compliant format. To make the raw metadata queryable, it is translated into a relational database specific schema (miniSQL) using the `sgml2sql` tool [11].

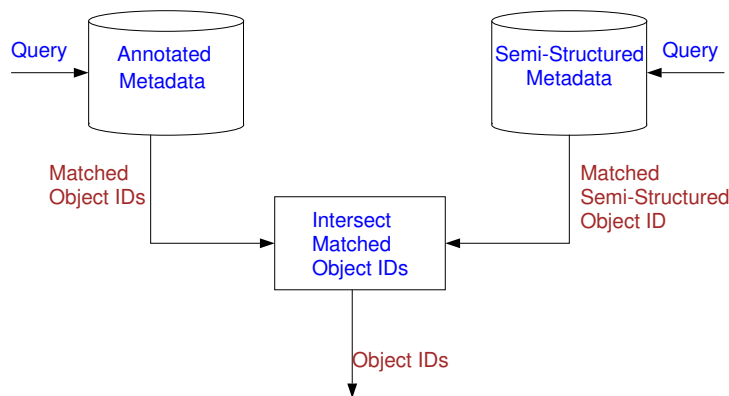


Figure 10: Process Diagram for Using Visual Metadata to Increase Precision

We decode the closed-caption data associated with video data, and convert them into unstructured metadata. The unstructured metadata are then indexed using SMART and the indices are stored in SMART-compliant files.

Queries are issued using the Web interface, which is implemented using HTML and the Java language. At the time of interface rendering, the annotated metadata are automatically extracted from the RDB and displayed. Queries composed by a user with the “point and click” method (annotated metadata-based query) are translated into SQL and sent to the relational database (miniSQL) for processing. A user can enter keywords (unstructured metadata-based query) that are converted into SMART-compliant query format for processing. A user can also simultaneously query both annotated and unstructured metadata (composite metadata-based query).

For the annotated metadata-based query, a Boolean matching technique is used to compare annotated metadata (manually extracted) and user specified criteria. If any segment belonging to a news item matches the query, then all the other segments belonging to the news item are retrieved based on the sibling relationship, and these segments form a candidate set. In an unstructured metadata-based query, the segments retrieved as a result of user specified criteria are clustered based on the similarity among the segments. Next, the clustered segments are augmented using a transitive search and the sibling relationships among the segments. The resulting clusters or candidate sets are used for compositions. In the prototype, we do not as yet take advantage of additional metadata that is available via audio (i.e., sounds that are not speech), we only use the spoken words in form of closed-caption data.

In the composite metadata-based query, the common segments retrieved from the two

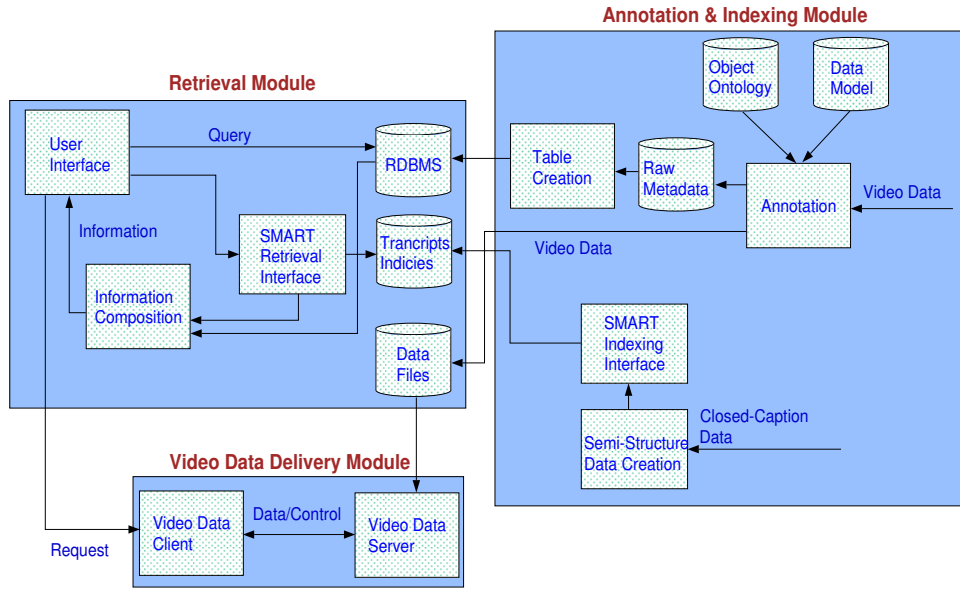


Figure 11: Architecture of the Digital Video Production System

individual queries (annotated metadata-based and unstructured metadata-based) are retained for composition.

The interface scripts used are written in the C language to execute the queries. The transitive retrieval technique and all the composition techniques are implemented as interface scripts. The conceptual compositions formed from the candidate sets are displayed in the Web interface, from where the user initiates video playback. Video data are streamed separately through the TCP/IP protocol and displayed using MTV, a MPEG-1 video playback client, or alternatively, using the Real Video player.

7 Conclusion

In this paper we propose a four-step hybrid retrieval technique that utilizes multiple metadata sets to isolate video information for composition. The technique relies on the availability of annotated metadata representing segment content and structure as well as segment transcripts that are unstructured. The unstructured metadata are readily available from closed-caption decoding or speech-to-text conversion tools. Basic content metadata such as date, time, and title can be recorded at the time of capture.

Our retrieval process applies a conventional approach to identifying segments using

content metadata. This is supported by clustering candidates into potential news items and a transitive search to increase recall. Finally, creation-time relationships expand the final candidate set of video segments. Experimental results on our data set indicate a significant increase in recall due to the use of additional concepts among related clips. We have demonstrated the results of the four step hybrid retrieval technique on unstructured metadata both across and within video clips, thereby increasing overall recall. Our retrieval technique also achieves clustering, which is necessary for creating cohesive video stories based on chronological or thematic ordering.

References

- [1] G. Ahanger and T.D.C Little, "The Canvass News Video Composition," On-line Demonstration, <http://hulk.bu.edu/projects/fap/demo.html>, 1999.
- [2] G. Ahanger and T.D.C Little, "Automatic Composition Techniques for Video Production," *IEEE Trans. on Knowledge and Data Engineering*, Vol. 10, No. 6, 1998, pp. 967-987.
- [3] G. Ahanger and T.D.C Little, "A Language to Support Automatic Composition of Newscasts," *Computing and Information Technology*, Vol. 6, No. 3, 1998, pp. 297-310.
- [4] G. Ahanger and T.D.C Little, "A System for Customized News Delivery from Video Archives" *Proc. Intl. Conf. on Multimedia Computing and Systems*, Ottawa, Canada, pp. 526-533, 1997.
- [5] G. Ahanger and T.D.C Little, "A Survey of Technologies for Parsing and Indexing Digital Video," *Visual Communication and Image Representation*, Vol. 7, No. 1, pp. 28-43, 1996.
- [6] A. Akutsu and Y. Tonomura, "Video Tomography; An Efficient Method for Camerawork Extraction and Motion Analysis," *Proc. ACM Multimedia '94*, San Francisco, CA, pp. 349-356, 1994.
- [7] E. Ardizzone and M. La Casia, "Automatic Video Database Indexing and Retrieval," *Multimedia Tools and Applications*, Vol. 4, No. 1, pp. 29-56, 1997.
- [8] E. Branigan, "Narrative Schema," in *Narrative Comprehension and Film*, pp. 1-32, Rutledge, New York, 1992.

- [9] M.G. Brown, J.T. Foote, G.J.F. Jones, K.S. Jones, and S.J. Young, "Automatic Content-Based Retrieval of Broadcast News," *Proc. ACM Multimedia '95*, San Francisco, CA, pages 35-43, 1995.
- [10] C. Buckley, Implementation of the SMART Information Retrieval System. Computer Science Department, Cornell University, No. TR85-686, 1985.
- [11] M. Carrer, L. Ligresti, G. Ahanger, and T.D.C. Little, "An Annotation Engine for Supporting Video Database Population," *Multimedia Tools and Applications* Vol. 5, No. 3, November 1997, pp. 233-258.
- [12] S.-F. Chang, J.R. Smith, M. Beigi, and A. Benitez, "Visual Information Retrieval from Large Distributed Online Repositories," *Communications of the ACM*, Vol. 40, No. 12, pp. 63-72, 1997.
- [13] T.H. Cormen, C.E. Leiserson, and R.L. Rivest, *Introduction to Algorithms*, MIT Press, 1990.
- [14] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 1. No. 2, April 1979.
- [15] R.O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.
- [16] J. Hafner, H. Sawney, W. Equitz, M. Flickner, and W. Niblack, "Efficient Color Histogram Indexing for Quadratic Form Distance Functions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 1, No. 7, pp. 729-736, 1995.
- [17] J.A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, 1975.
- [18] W. Klippgen, T.D.C. Little, G. Ahanger, and D. Venkatesh, "The Use of Metadata for the Rendering of Personalized Video Delivery," In Amit Sheth and Wolfgang Klas, eds., *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*, pp. 287-318, McGraw Hill, New York, 1998.
- [19] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Video Abstracting," *Communications of the ACM*, Vol. 40, No. 12, pp. 55-62, 1997.
- [20] V.E. Ogle and M. Stonebreaker, "Chabot: Retrieval from a Relational Database of Images," *Computer*, 28(2):49-56, 1995.

- [21] R. Picard and T. Minka, "Vision Texture for Annotation," *Multimedia Systems*, 3(3):3-14, 1995.
- [22] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, New York, 1983.
- [23] S. Santini and R. Jain, "Similarity is a Geometer," *Multimedia Tools and Applications*, Vol. 5, No. 3, pp. 277-306, 1997.
- [24] S. Sclaroff and J. Isidoro, "Active Blobs," *Proc. Intl. Conf. on Computer Vision*, Mumbai, India, 1998.
- [25] J.S. Wachman, "A Video Browser that Learns by Example," *Master Thesis*, Technical Report #383, MIT Media Laboratory, Cambridge, MA, 1997.
- [26] H. Wactlar, T. Kanade, M.A. Smith, and S.M. Stevens, "Intelligent Access to Digital Video: The Informedia Project," *Computer*, 29(5):46-52, 1996.

Table 5: Sample Unstructured Metadata

.idDoc: cnn2.txt/O193 .videoFile: d65.mps .textData: Justice correspondent Pierre Thomas looks at the long-awaited decision. After months of intense pressure, attorney general Janet Reno has made a series of decisions sure to ignite a new round of political warfare. Regarding fund raising telephone calls by Mr. Clinton at the White House: no independent counsel. On vice president Gore's fund raising calls: no independent counsel. Controversial democratic campaign fund-raiser Johnny Chung has alleged he donated 25,000 to O'Leary's favorite charity in exchange for a meeting between O'Leary and a Chinese business associate. Three calls for an independent counsel. All three rejected.

Table 6: Weight Assignment

Doc ID	Concept	Scheme 1	Scheme 2
146	barred	0.62630	4.04180
146	weapons	0.15533	2.50603
146	iraqi	0.21202	
146	u.n	0.18075	2.72990
146	continues	0.31821	2.58237
146	standoff	0.36409	3.87444
146	iraq	0.13211	2.71492
146	sights	0.50471	4.04180

Table 7: System Performance

Search Technique	Recall	Precision
Query Match	29%	100%
Transitive Search	42%	90%
Sibling Identification	88%	85%