Matching Songs to Events in Image Collections

Mark D. Wood
Eastman Kodak Company
mark.d.wood@kodak.com

Abstract

This work identifies relevant songs from a user's personal music collection to accompany pictures of an event. The event's pictures are analyzed to extract aggregated semantic concepts in a variety of dimensions, including scene type, geospatial information, and event type, along with user-provided keywords. These semantic concepts are then used to form a search query against a song database based primarily on the song lyrics. Songs are scored using probabilistic techniques to come up with a rank ordered list of candidate songs that could then be used as, e.g., the audio track in a multimedia slideshow.

1. Introduction

Multimedia presentations composed of still imagery and video content benefit greatly from an accompanying audio track. In this era of digital communication and social networking, people increasingly want to quickly create and share multimedia with their social network. The multimedia content may serve to memorialize a person or document a significant event, such as a 50th anniversary or a baby's first steps, or it may simply serve a more transitory purpose, such as to entertain or inform. Adding music to the still or video content further engages the senses; appropriately selected content enhances the mood or message the composer wishes to convey. However, choosing appropriate content is often a mind-numbing task, as the amount of candidate content can easily number into the thousands. This research provides tools to assist people in selecting appropriate songs[1] to accompany their multimedia creations.

This work suggests songs by first conducting a semantic analysis of the source content—the pictures and video snippets that form the basis of the multimedia composition. The system then uses these semantic concepts as search terms against a music database to generate a list of semantically relevant songs from which the user can choose one or more to use as the audio track. The music database is indexed primarily by the lyrics of the songs, although additional metadata such as the song genre can also play into the matching process.

The next section briefly describes other efforts at matching music with pictures and video. Section 3 presents the basic system approach, and gives an overview of the Indri search engine used in this work. Section 4 describes how semantic concepts are used to form search queries. Section 5 presents some preliminary results from using the system. Finally, Section 6 summarizes the work and points to future research opportunities.

2. Related Work

Previous work has considered limited aspects of the problem addressed here. In [1], the authors extract semantic concepts from the lyrics of songs and generate search terms from those concepts to identify appropriate pictures from online sources to go with the lyrics, in order to produce a music video. The work of [14] produces slide shows of personal photos to accompany a given song's lyrics, where no semantic information is assumed to be associated with the personal photos. Instead, image analysis techniques are used to determine similarity between personal photos and reference photos found online using lyric phrases with Google™ Image Search. In [8], a multimedia authoring system uses scene classifiers to classify a set of pictures, where the scene class is then mapped to a specific music genre. This work is differentiated by its goal of characterizing picture-taking events as a set of multidimensional semantic

[1] Disclaimer: Music and other creative works are protected by copyright law. Permission of the copyright owner may be required before such material can be incorporated into multimedia presentations and other uses.

978-0-7695-3800-6/09 $26.00 © 2009 IEEE
DOI 10.1109/ICSC.2009.33

concepts, which are then used to form search terms against a music database.

3. System Design

The system consists of several major functional components as illustrated in Figure 1, including components for semantically indexing assets, a database for storing semantic information, and a database containing song information. The semantic indexing is broken up into two parts. The first part generates semantic information for each asset. Such information may include scene and material classifiers [7][11], age and gender classifiers, and geospatial classifiers. It also includes extracting basic information directly from the file metadata, such as capture time and user-supplied keywords. In addition, it includes algorithms for detecting and classifying events [2]. In this work, an event simply defines a set of input images, and may consist of one or more levels of subevents. The generated metadata is stored in a metadata database; we used the Franz AllegroGraph triplestore product [3] as it supports a rich query language with inferencing capabilities. These types of metadata are precomputed and stored in the triplestore. A separate, second indexing component further generates summarizing semantic information for selected events on an as-needed basis.

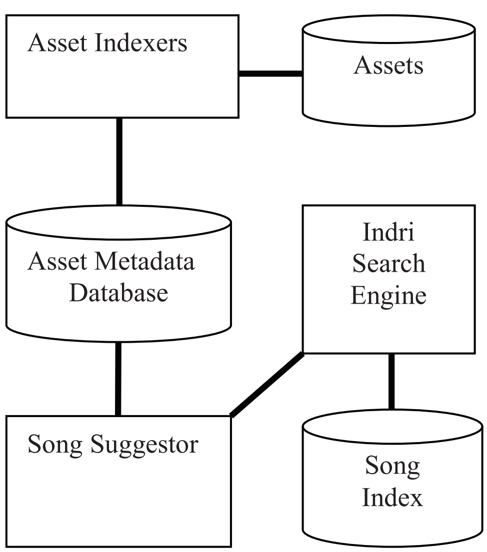

Figure 1. System Components

In addition to asset metadata, the system includes an indexed database of songs. In this work, the song database is populated by querying Microsoft® Windows Media Player to get the list of songs, with the title, artist, and genre. By using personal content, the system is more likely to choose songs that the user likes. The system then attempts to find the associated lyrics for each song using the webservice provided by LyricWiki [4]. This community-driven website provides lyrics for many popular and contemporary songs. Commercial service providers such as Gracenote® provide similar services on a per-fee basis. A separate XML-based document is constructed containing the lyrics of each song, along with the title, artist, and genre. The collection of documents is then indexed using the Indri information retrieval system, described in more detail in Section 3.1.

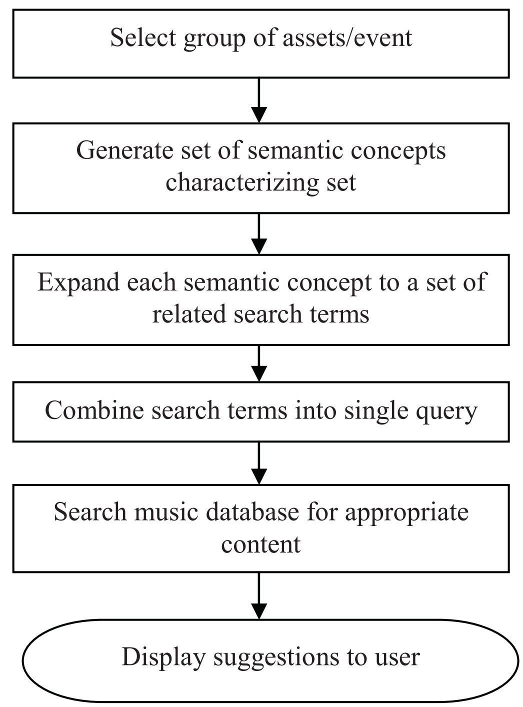

Figure 2: System Workflow

Figure 2 summarizes the basic system workflow. First, the user or system identifies a specific event for which a song should be selected. The event (or set of events) currently is manually selected by the user. Alternatively, the system could employ some automated means for asset selection in order to produce a story or event summary, such as the method described in [13]. An event should correspond to a set of semantically related pictures, such as a party, an outing, or a holiday. Some types of events, such as vacations, may span a wide variety of subevents. In such cases, the system will attempt to find a song that best characterizes the event as a whole. However, it may be more pleasing to the user to split the event up into subevents, and find separate songs for each subevent. The song suggestor generates a series of queries to the asset metadata database in order to produce a set of semantic concepts characterizing the target event. Each concept is further expanded using a concept expander. Based on this semantic information, the song suggestor then generates a search query. The search query consists of a set of weighted search terms. The song suggestor passes the query to the Indri query engine to evaluate, getting back a list of candidate songs with associated scores.

3.1. Song Lyric Indexing

In order to match semantic concepts to songs, we needed a mechanism for searching through song lyrics and other metadata to compute a probabilistic score for each song. The Indri search engine [12][9] uses a probabilistic inference network framework in order to score documents against complex queries. For a given query, Indri constructs an inference network, which is used to estimate the probability that a given document satisfies the query. This information retrieval model matched the need of this work very well, and so Indri was used as the search engine for querying a music repository.

Indri constructs a multiple Bernoulli model [10] for each section of a song, where in our work the sections include the title, the lyrics, and the genre. Indri supports arbitrarily complex queries consisting of series of terms, where terms include single words, ordered word phrases, and (optionally) weighted synonyms. Terms may be combined using various weighted and unweighted belief operators, as well as a negation operator. Figure 4 illustrates the basic structure of an inference network applied to a song document. Indri forms a feature language model for each of the separate parts of the XML document. A given Indri query consists of one or more levels of belief operators, such as #weigh and #combine, which combine various beliefs in a weighted or unweighted manner. Belief nodes may combine probabilities from other belief nodes, or from representation nodes. Representation nodes represent the likelihood of a given feature occurring in a given document. Features are represented in the Indri query language as any of several types of terms. The most basic term is simply a word, but Indri also supports phrase and proximity terms, as well as terms to indicate synonyms.

4. Query Generation

Once the set of input images has been analyzed to extract semantic concepts, the system then generates an Indri query, which consists of a set of weighted search terms. More precisely, each search term here represents an Indri belief operator expression, as defined in the Indri literature. The query is currently formed by considering the following key semantic concepts:

- holidays
- event type
- scene and materials types
- geospatial information
- user-provided keywords

The system runs semantic indexers for each of these categories. The algorithms typically return a set of labels, with optional belief values or scores. The system uses the concept expander to expand each concept into a set of typically weighted search terms. The concept expander takes words or phrases such as "Christmas" or "lake," filters out stopwords, and then returns a valid Indri query expression that includes corresponding terms likely to be found in songs. For example, the term "lake" is expanded to the weighted list of synonyms #wsyn(1.0 lake 0.8 pond 0.7 lagoon 0.7 waterhole 0.5 water). In the current version of the system the expansions were manually determined, particularly for concepts returned by the classifiers, by consulting appropriate sources such as WordNet® and common songs. If the concept expander does not have an expansion for a particular word, it simply returns the word itself. More robust concept expansion could be automatically carried out using sources such as WordNet to find synonyms. Finally, a single query is formed as a weighted combination of each of these subquery expressions.

4.1. Holiday Classification

The date of a picture-taking activity can in many cases provide a clue as to the picture content. For example, pictures taken on or near Christmas are likely to be related to Christmas. For each asset in the event, the system attempts to determine if the day is near—within three days—of a known holiday such as Christmas or Valentine's Day. For each such holiday, a score is assigned to the picture, based on how close the capture date is to the holiday, using an exponential backoff weighting function. If the picture was taken on the holiday, the score is 1.0, if the picture is taken within a day of the holiday, the score is 0.5, and so on. To form a set of aggregate scores, the system sums the scores for each holiday and divides by the number of assets. So if all the pictures were taken on Christmas, the aggregate score for Christmas would be 1.0; if half were taken on Christmas and half were taken the day after Christmas, then the aggregate score would be 0.75. This score is used as the weight for the query.

The three-day window around holidays was arbitrarily chosen. A picture taken one week before Christmas is expected to be more likely related to Christmas than a picture taken one week before

Valentine's Day is likely to be related to Valentine's Day. The model could easily be extended to accommodate such differences, given research relating picture-taking activity to specific holidays.

Given a set of one or more holidays with associated scores, the holiday portion of the query is formed by expanding the name of the holiday into appropriate concepts using the concept expander. In the current version, the concepts associated with a holiday were manually chosen. However, they could be more systematically chosen by analyzing songs associated with those holidays, using lists such as those maintained at www.popculturemadness.com.

4.2. Event Classification

The event classification algorithm of [2] is used to characterize the set of pictures as belonging to one of four types: family moments, party, outdoor sports, or vacation. This classification is already done at the event level, and so no further aggregation is done.

The event classification algorithm uses a Bayes net to compute a probability associated with each class. Unfortunately, the version of the algorithm used here simply returned the highest scoring label without providing the score. The returned label is mapped through the concept expander, which has specially constructed expansions for the four terms used by the event classifier.

4.3. Scene Classification

We considered several different scene classifiers. The one [8] used here produces a rather limited number of outdoor scene class types (beach, urban, mountain, fall foliage, field, sunset). Other scene classifiers work on indoor scenes, returning classifications such as "family room" or "bedroom." However such classifications seem less useful in choosing a song than the outdoor scene classifiers.

The classifier is run on a given image only if the probability of the image representing an outdoor scene is sufficiently high. The classifier returns a score for each image, not a true probability. In order to compute a scene classification for a group of images, the system takes all the scores for each class, throws out the highest score, and then computes the average of the next three highest scores. This approach was shown in [8] to provide a good characterization of the grouping without being unduly influenced by a single image. However, if the number of images in the group is less than four, then the system simply computes the average score. The scores are then normalized to a value between 0 and 1.0 by using a sigmoid function. The normalized value is used as the weight for the scene part of the query in the final query. To further limit the impact of a few images unduly influencing the event summary, the scene information is only included in the final query if a sufficient percentage of the images have a valid scene classification. For example, if a group of images was predominantly shot indoors, but one or two pictures were shot outdoors, the outdoor scene classification may not be relevant.

Since the chosen outdoor scene classifiers are rather limited, they are supplemented by the use of some material classifiers, specifically, snow and water. When applied to a given image, these classifiers return a measure of what proportion of the image is believed to be covered with that material, such as water or snow. To aggregate these values, the system queries for each material classification of interest, but only considers the material classification for a given image if the probability that the image is of an outdoor scene is greater than 0.5. The aggregate material classification is computed as the average of the individual image material classification values. The system also records for each material class how many images were counted as having that class. A material class is only considered in generating the final query if a sufficient percentage of the images had that material classification, again to prevent one or two images from unduly impacting the result.

4.4. Geospatial Classification

The geospatial indexer was developed specifically for this work. It takes available latitude and longitude information and uses a reverse geocoder to identify nearby feature points and a set of possible feature class types describing the location. It also obtains the nearest city or town name, and the associated state/province and country.

The publicly available geonames.org database [5] is formed from a collection of national databases supplemented with user-provided content. It contains a list of place names, their associated feature type (such as mountain, school, lake, etc.), and the corresponding latitude/longitude for the approximate centroid for the feature point. Unfortunately, this database does not provide for each feature point a definition of the geographic region encompassed by the feature. A given image latitude, longitude pair is unlikely to map exactly to one of the geonames.org feature points. Consequently, the geospatial indexer asks for the twenty closest matches within a three-kilometer radius. Some of these feature points may be relevant; many or most may not be relevant. As a heuristic, the geospatial indexer takes the up to twenty feature points in order of their distance from the target latitude, longitude, and computes the average delta in the feature point distances from the target as well as the standard deviation. The geospatial indexer then computes a cutoff point by starting at the feature point closest to the target point, and going outward, stopping at a point where the delta in distance exceeds the average delta plus the standard deviation. Clearly this is simple heuristic; other work [6] attempts to compute 2-means clusters. Figure 3 illustrates a situation where this heuristic works well: the target point, represented by a diamond, is nearby some relevant points of interest (Homosassa Springs and Homosassa Wildlife State Park), while the other feature points are clearly separated away from the target point.

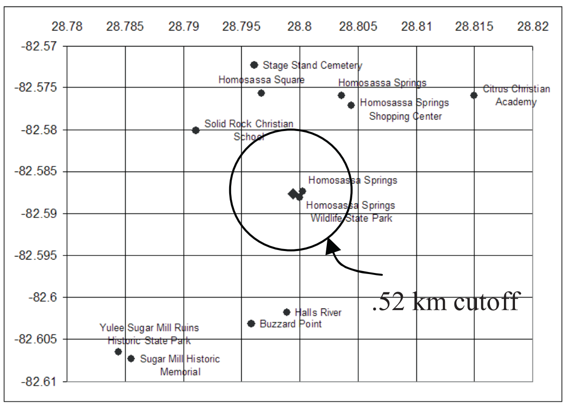

Figure 3. Geospatial Example

Once the set of nearby feature points has been identified, the geospatial indexer then computes a set of possible feature classes describing the target. The feature classes are determined by taking a weighted average of the feature classes associated with each of the nearby feature points. If no feature point is within the cutoff distance, then the geospatial indexer examines the feature class type of the nearest feature point within the three-kilometer radius. If that type is a type typically denoting a large area (such as mountain, lake or park) as opposed to a single point (such as a school), then that feature class is assumed to characterize the target. This heuristic could be further refined. For example, relatively nearby feature points whose feature type is associated with a relatively small area (e.g., a building) could be ignored in favor of a more distant feature point whose feature type typically denotes a large area (such as a park).

In addition to examining the feature class types, the geospatial analyzer also produces a bag of weighted keywords, similar to [6]. The list of keywords is formed from the words used in the placenames, again weighted according to the distance of the associated feature point from the target.

4.5. User-Provided Keywords

When available, user-provided keywords can be one of the best sources of semantic information. Such keywords are directly mapped by the concept expander to a set of search terms. Although not currently implemented, a mechanism such as WordNet could be readily employed to map keywords to synonym sets.

4.6. Relative Weighting

To produce the final search query for the Indri query engine, the previous search terms are combined into a single belief expression using experimentally tested weights. The weights for each term are computed as follows:

- The weight for each holiday search term is the score, computed as described in Section 4.1.
- Scene-based search terms are weighted by normalizing the score for each scene type using a sigmoid function and the material scores are used as the weights, as described in Section 4.3.
- Geospatial feature classes and keywords are assigned the weight of 0.9; the state/place information is assigned a weight of 0.3.
- The event classification is assigned a weight of 0.7, given the current absence of a corresponding score.
- Search terms based on user-defined keywords are given the highest weight, 2.0.

5. Results

The system has been tested against several different music collections ranging in size from around a hundred songs to several hundred songs. The system imposes no inherent limit on either the number of pictures nor the number of songs in the music collection. For the purposes of demonstration here, the system was tested against a small song database containing 106 songs from artists such as Enya, Johnny Cash, and Gordon Lightfoot. These songs were indexed using Indri version 2.8, with common stopwords filtered out as part of the indexing. Figure 5 illustrates a usage of the system on an event containing six pictures taken February 15, 2005 on Pine Island, Florida. The left portion of the frame shows thumbnails of the images in the event. The right side of the frame shows the extracted semantic concepts, the generated Indri query, and the resulting list of songs. The example is small for illustrative purposes; larger event sets tend to produce even more semantic information, which results in longer, more detailed queries.

In the list of semantic concepts, we see that the system correctly classified this event as “vacation,” and determined that the event was near Valentine’s Day. Furthermore the system rightly concluded with high probability that the event took place outdoors, and at a beach, with significant water present. For this particular set of pictures, no user-defined keywords were present, and so all the semantic information was gleaned using analytical means. One or more pictures in the set were geotagged, however, and this resulted in the feature type of “island” being associated with the event, as well as the set of keywords extracted from the nearby placenames.

Below the list of semantic concepts is the generated Indri query. The outermost #wsum operator [9] is used to combine two belief expressions: the primary belief expression implied by the extracted semantic concepts, and a belief expression used to deemphasize more “negative” songs—songs with more morbid themes—under the assumption that such songs are less likely to be of interest to the user as a musical accompaniment. However, note that such songs may still be returned if the primary belief expression scores such songs highly. The primary belief expression contains elements corresponding to the various semantic concepts. Since the event occurred near Valentine’s Day, the query includes keywords associated with romantic elements, but the weight for this term is only 0.5, since the pictures were taken the day after Valentine’s Day. The beach, water, and geospatial concepts are all expanded by the concept expander to appropriate search terms.

The lower right corner contains the list of songs returned by executing the Indri query, with their associated log-value scores. For the purposes of this work, Indri was run using its default parameters, including the use of Dirichlet smoothing with $\mu = 2500$. The island, beach, sand, and Florida concepts are all reflected in the list of returned songs. The resulting list of songs should be treated, of course, as simply a suggested list of relevant songs, with some more likely to be deemed relevant by the user than others. In this case, the Enya “Orinoco Flow” and “On Your Shore” pieces, along with the Johnny Cash “Orange Blossom Express” might be considered reasonable candidates. Songs such as “Whispering Pines” and “Bonanza” were returned because they reference pine trees, and therefore match the placename Pine Island; however they clearly are not relevant to the picture matter.

6. Discussion

The goal of this work is to help users choose suitable songs to accompany the pictures of an event. The approach is successful if at least one suggested song meets the user’s approval; we do not expect the user to find every suggested song to be a perfect match. While early results from this work show promise, they also highlight some of the difficulties associated with matching semantic concepts to songs. Existing scene and activity classifiers have limited accuracy, and the types of classifications they return are not always helpful in mapping to appropriate music.

Additional semantic information directly provided by the user in the form of captions or keywords may be more meaningful than algorithmically derived information. However, people seldom take the trouble of annotating their picture collections. Annotations are perhaps most commonly made in the context of picture sharing through sites such as Facebook, but then such comments may be less descriptive in nature. Another possible source of information is a user’s calendar. By combining multiple sources of semantic information, we are able to produce a richer characterization of an event than would be available from any one source.

An obvious extension would be to use inferencing techniques combined with auxiliary sources of information to fuse semantic information from different classifiers to provide an even more accurate classification. For example, the event type and geospatial feature type might be combined to provide a more descriptive event type classification, which could be used to improve the accuracy of the music matches. A set of vacation pictures in a park might be more likely to be a camping activity, whereas a set of vacation pictures in an urban location might indicate some type of sight-seeing. In Figure 5, semantic fusing could have been used to deemphasize the concept of “pines,” since the pictured subject matter did not include a significant number of pine trees. Knowledge of personal user preferences and interests can also be factored into weights. Further work is required both to determine the optimal weighting of the different classes of search terms as well as the interdependencies between the different classes.

We have considered using other sources of data such as age and gender classifiers, as well as a more extensive set of scene classifiers. While the performance of such classifiers did not yet seem sufficiently robust enough to include now, new classifiers can be readily included in the future. As additional image classifiers become available, they can provide further types of semantic information. For example, the current query generator does not add constraints with respect to genre. However, some sources of semantic information might suggest certain musical genres. For example, people age classifiers might enable the system to determine if an event primarily features young children; this could cause the system to give preference toward children's music.

This work currently characterizes songs solely based on textual analysis. Songs are a form of poetry. As such, they typically make use of metaphors and similes, and such constructs can result in artificial matches. For example, the country song "Sink the Bismarck" describing the sinking of a German WWII battleship includes repeated references to "guns as big as trees," but that hardly qualifies the song as suitable accompaniment for pictures of foliage! However, one can just as easily come up with other examples where the use of a simile or metaphor would result in an appropriate match.

Textual analysis misses other aspects of music, such as the tempo, which should also feed into the selection process. For example, pictures of an active sporting event typically should be accompanied by more lively music. Moreover, this text-based approach does not work for instrumental music lacking a descriptive title. Future work could combine textual analysis with work that automatically characterizes music by using signal processing techniques. And of course music characterizations may be done manually, where people describe the type of event for which the music is appropriate. Such music annotations could be generated using community-generated information and analysis, in a wiki-like fashion. Regardless of how the music characterization is done, this work provides a means for extracting and matching event-based semantic concepts to the music.

The success of this approach hinges on accurate semantic understanding of both events and songs. Initial results look promising, but illustrate the need for better classifiers. Leveraging additional contextual information, such as a user's personal calendar, will likely further improve the event classification. Likewise, the lyric-based indexing of songs used in this work is satisfactory for more literal songs, but may not capture the spirit of more poetic lyrics. Even with these limitations, this work provides a basis for suggesting music to accompany user imagery; further classifier improvements will translate into direct improvements in song suggestions.

7. Acknowledgments

Jiebo Luo provided helpful feedback and input into this work; thanks also to the anonymous reviewers for their constructive feedback.

8. References

[1] Cai, Rui, Lei Zhang, Feng Jing, Wei Lai and Wei-Ying Ma, "Automatic Music Video Generation Using Web Image Resource," IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, 2007.

[2] Das, Madirakshi and Alexander Loui, "Event Classification in Personal Image Collections," ICME Workshop on Media Information Analysis for Personal and Social Applications, July 2009.

[3] Franz, Inc. AllegroGraph RDFStore. http://www.franz.com/agraph/allegrograph/.

[4] http://lyricwiki.org/.

[5] http://www.geonames.org.

[6] Joshi, Dhiraj and Jiebo Luo, "Inferring Generic Activities and Events from Image Content and Bags of Geo-tags," ACM Intl. Conf. on Image and Video Retrieval, Niagara Falls, Canada, July 2008.

[7] Luo, Jiebo and Matthew Boutell, "Natural Scene Classification Using Overcomplete ICA," *Pattern Recognition*, vol. 38, pp. 1507-1519, 2005.

[8] Luo, Jiebo, Alexander Loui, Matthew Boutell and Phoury Lei, "Photo-Centric Multimedia Authoring Enhanced by Cross-Media Indexing," SPIE Intl. Symposium on Visual Communication and Image Processing, Special Session on Digital Entertainment, Beijing, China, July 2005.

[9] Metzler, Donald, "Indri Retrieval Model Overview," http://ciir.cs.umass.edu/~metzler/indriretmodel.html.

[10] Metzler, Donald, Victor Lavrenko and W. Bruce Croft, "Formal Multiple-Bernoulli Models for Language Modeling," Proceedings of ACM SIGIR 2004, 540-541, Sheffield, England, 2004.

[11] Singhal, Amit, Jiebo Luo and Weiyu Zhu. "Probabilistic Spatial Context Models for Scene Content Understanding," IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Madison, Wisconsin. 2003.

[12] Strohman, Trevor, Donald Metzler, Howard Turtle and W. Bruce Croft, "Indri: A language-model based search engine for complex queries," University of Massachusetts-Amherst CIIR Technical Report, 2005.

[13] Wood, Mark D., "Exploiting Semantics for Personalized Story Creation," IEEE Intl. Conf. on Semantic Computing, Santa Clara, CA, August 2008.

[14] Xu, Songhua, Tao Jin and Francis C. M. Lau, "Automatic Generation of Music Slide Show using Personal Photos," 10th IEEE Intl. Symposium on Multimedia, December 2008.

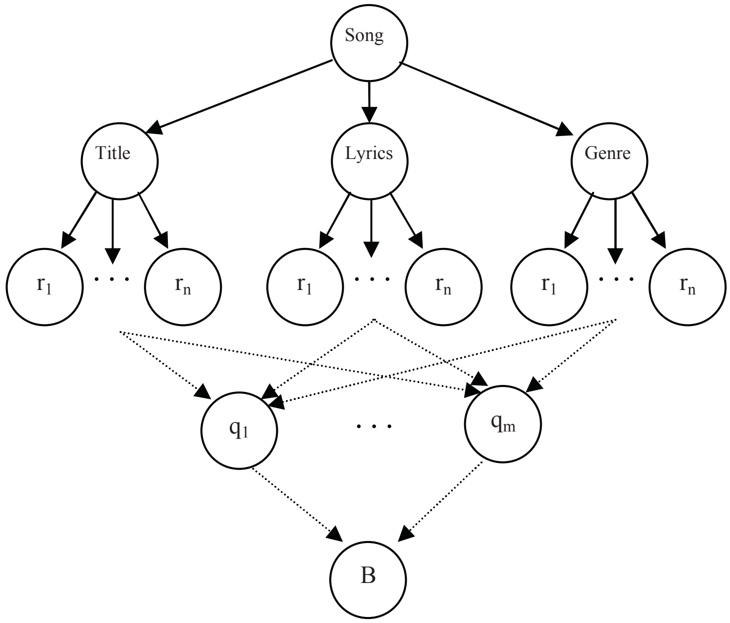

Figure 4. Inference Network

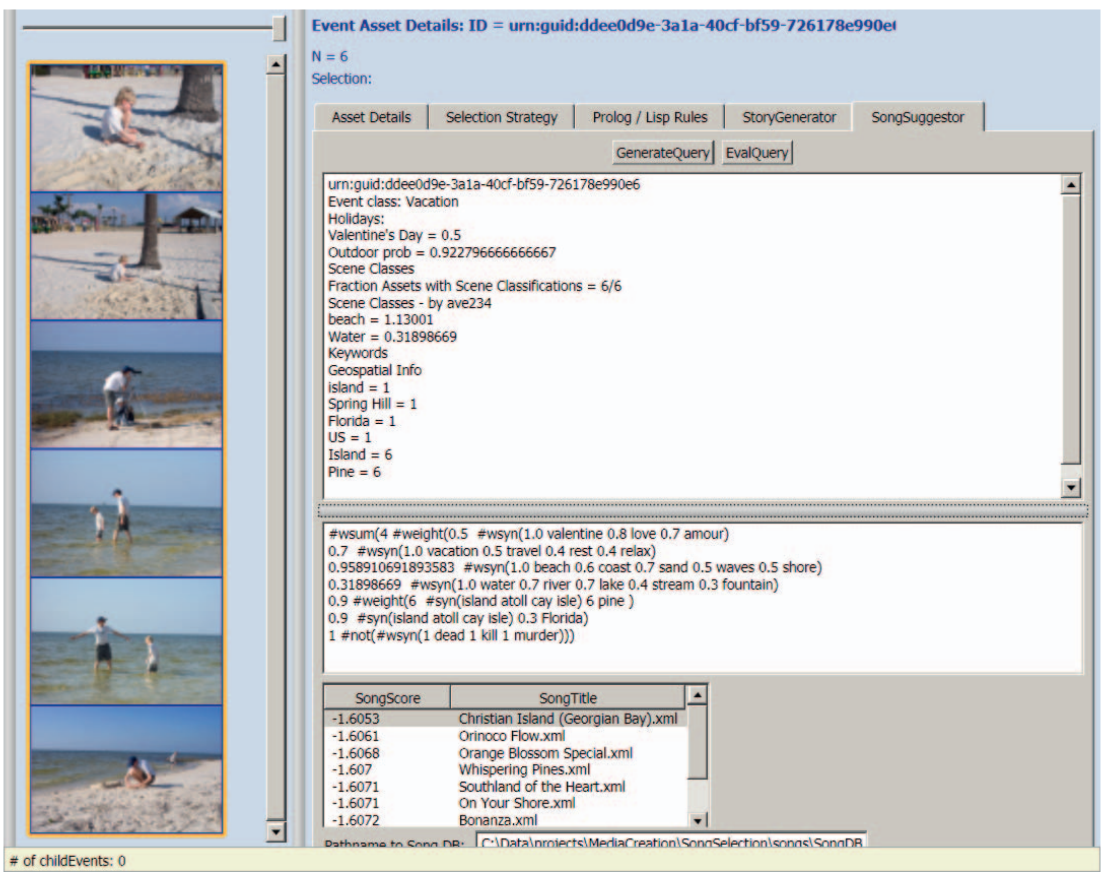

Figure 5. Example Results