Video Search Re-Ranking via Multi-Graph Propagation

Jingjing Liu
Department of Computer Science
Nankai University
Tianjin, China
nkicestone@gmail.com

Wei Lai
Microsoft Research Asia
No. 49 Zhichun Rd.
Beijing, China
weilai@microsoft.com

Xian-Sheng Hua
Microsoft Research Asia
No. 49 Zhichun Rd.
Beijing, China
xshua@microsoft.com

Yalou Huang
College of Software
Nankai University
Tianjin, China
huangyl@nankai.edu.cn

Shipeng Li
Microsoft Research Asia
No. 49 Zhichun Rd.
Beijing, China
spli@microsoft.com

ABSTRACT

This paper[1] is concerned with the problem of multimodal fusion in video search. First, we employ an *object-sensitive* approach to query analysis to improve the baseline result of text-based video search. Then, we propose a *PageRank-like* graph-based approach to text-based search result re-ranking. To better exploit the underlying relationship between video shots, the proposed re-ranking scheme simultaneously leverages textual relevancy, semantic concept relevancy, and low-level-feature-based visual similarity. In this PageRank-like scheme, we construct a set of graphs with the video shots as vertexes, and the conceptual and visual similarity between video shots as "hyperlinks." A modified topic-sensitive PageRank algorithm is then applied on these graphs to propagate the relevance scores through all related video shots. Experimental results verify the effectiveness of the graph-based propagation approach combined with the object-sensitive query analysis approach, which brings significant improvement to the baseline of text-based video search. Our experimental analysis also indicates that the proposed re-ranking method is highly generic and independent of different query classes, training data, and human interference.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms

Algorithms, Performance, Experimentation

Keywords

Video search, multimodal fusion, PageRank algorithm, multi-graph propagation, re-ranking, object-sensitive, query analysis

1. INTRODUCTION

There is a rapid growth of online video data as well as personal video recordings in our daily life. In order to successfully manage and use such enormous multimedia resources, users need to be able to conduct semantic searches over the multimodal corpora efficiently and effectively. This leads to a continuingly growing interest in video retrieval in the research community recently.

Video search is an active and challenging task. It is defined as searching for the relevant video or video segments/clips with issued textual queries (keywords, phrases, or sentences) and/or provided video clips or image examples (or some combination of the two). Many approaches have been tested in recent years, ranging from plainly associating video shots with text search scores to sophisticated fusion of multiple modalities [1][9][14][17][19][21]. It has been proved that the additional use of other available modalities such as image content, audio, face detection, and high-level semantic concept detection can effectively improve pure text-based video search.

A generic hierarchical framework of video search system is illustrated in Fig. 1. A typical video search system consists of several main components: query analysis, uni-modal search, and re-ranking through multimodal fusion. By analyzing the given query with multiple types of information, different forms of the query (text, image, video, etc.) are input to individual search models, such as text-based search model, query by example (QBE) model and concept detection model. Then a fusion model is applied to aggregate the search results of the multimodalities.

Such video retrieval systems tend to get the most improvement in a multimodal fusion fashion by leveraging text search engines, multiple query example images, and specific semantic concept detectors. However, applying a universal fusion model independent of queries (e.g. the weight for each uni-modal is fixed for all queries in the fusion system) will lead to much noise and inaccuracy. Since this kind of retrieval by leveraging multimodalities across various textual and visual information sources, though promising, strongly depends on the characteristics of the specified queries. Therefore, in most multimodal fusion systems for video search, different fusion models are constructed for different query classes [4][5][6][12][13], with the involvement

[1] This work was performed when Jingjing Liu was visiting Microsoft Research Asia as a research intern.

MM'07, September 23-28, 2007, Augsburg, Bavaria, Germany.
Copyright 2007 ACM 978-1-59593-701-8/07/0009...$5.00.

of human knowledge. However, this laboratory-style fusion method is dependent on the quantity of human interference as well as the quality of employed human intelligence, which lacks in the adaptability to generic types of queries.

Many researchers have studied on how to better fuse the multimodalities in video search [7][18][19][21][23]. Kennedy et al. [23] proposed a query classification method to automatically discover the classes of query for a query-class-dependent multimodal fusion. By training a query classification model, the proposed framework learns the best linear fusion weights of uni-modals for each specific query class. Although this approach automatically discovers query classes, it is difficult to develop highly fine-tuned models for every class of queries; and the system cannot be easily ported to new domains or data sets which contain unknown classes of query. Moreover, the supervised learning process of query clusters is a data-driven method and requires much training data; therefore, it is not practical for large-scale video retrieval systems. It is clear that we need to explore automatic fusion approaches which can automatically adapt to and leverage the available textual and multimedia cues for video search, independent of specific queries, training data, and human knowledge.

Hsu et al. [19] proposed an IB-based (information bottleneck) re-ranking scheme for video/image search, which reorders results from text-only searches by discovering the salient visual patterns of relevant and irrelevant shots from the approximate relevance provided by text results. This approach, although leverages the low-level visual features of video shots with text search baseline, lacks the consideration of high-level conceptual relation between the video shots. Video shots with much visual similarity in low-level features may have no resemblance in high level features (conceptual level) due to the "semantic gap" between visual features and conceptual features. To by-pass the semantic gap, we need a finer way to smoothly leverage both conceptual and visual information into video search re-ranking.

Enlightened by this observation, in this paper we investigate on automatic multimodal fusion for video search, by employing not only textual and visual features, but also semantic and conceptual similarity between video shots to re-rank the search results. We aim to develop an approach to video search which not only can avoid the dependency on specific query characteristics, training data and human interference, but also can leverage textual relevancy, semantic concept relevancy, and visual similarity in a novel fashion. It would smooth the multimodal information sources in an implicit yet "soft" graph-based propagation way instead of an explicit and "hard" linear aggregation.

In the research area of web page retrieval, many studies have considered the union of text and graph-based link analysis. Nie et al. constructed a topical link analysis for web search [30]. Kurland and Lee studied on structural web page re-ranking utilizing cluster-based language models [24][25]. There have also been some studies on transferring graph-based approaches from text retrieval into multimedia processing in recent years. Shipman et al. [36][37][38] developed the concept of "detail-on-demand" video for "hypervideo" authoring where navigational links can be created between any two video clips or composites. The "hyperlinks" among video clips or composites construct a hierarchical navigation structure of video segments, similar to that of web pages with hyperlinks. Wang et al. [44] proposed to use random walk with restarts to implement image annotation refinement. To re-rank the selected annotations of images, a graph-based algorithm using Random Walk with Restarts (RWR) is proposed to leverage both the corpus information and the original confidence information of the annotations.

A typical random walk method for web page processing through hyperlinks is the PageRank algorithm [31], which is widely used in web page retrieval tasks. An assumption in the PageRank algorithm is that: the hyperlinks between web pages indicate the relative importance of web pages. The more hyperlinks point to a web page, the more important this web page is. In the original PageRank algorithm [31], a single PageRank vector is computed to capture the relative importance of web pages, using the link structure of the web independent of any particular search query. In [15], Haveliwala proposed a topic-sensitive PageRank approach biased towards a set of representative topics, to capture more accurately the notion of importance with respect to a particular topic. In the topic-sensitive PageRank, the more relevance to the given topic a web page has, the more important the web page is.

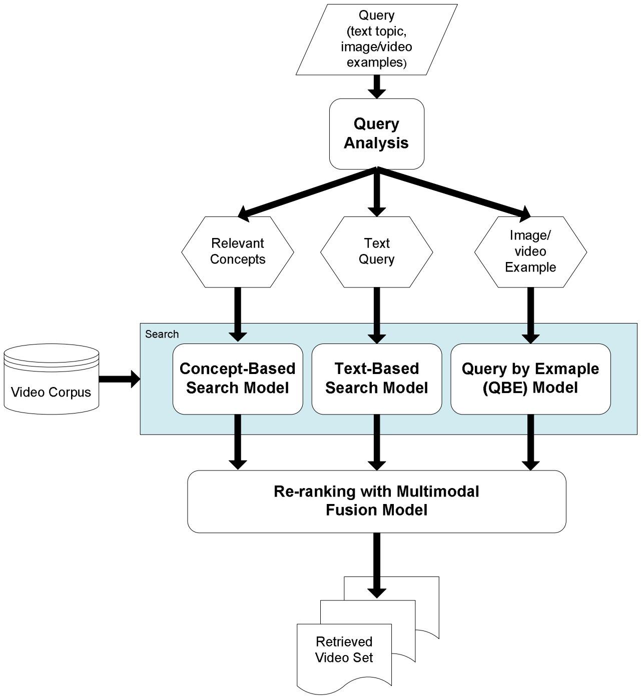

Fig. 1. Framework of the proposed video search system.

We observe that a similar relative relevance dependent on the given topic also exists in the video search tasks. In video corpus, each video clip is annotated with a set of semantic concepts, which represent the semantic content of the video clip. Therefore, given a query topic in text, the video clip whose concept labels are similar to the given topic is more likely to be relevant to the query. This is similar to the relevance of web pages to a given topic in web search tasks. Moreover, video shots are not independent of each other, but have mutual relations such as conceptual and visual similarity. This can be taken as the underlying "hyperlink" between video shots, similar to that between web pages. Therefore, an intuition is that, by adopting a topic-sensitive web page ranking algorithm into video search, the relevance of video shots to a given query can be learned from these "hyperlinks" indicating conceptual and visual relations similarly, which will improve the ranking results from pure text-based search model.

Based on this observation, we propose a PageRank-like approach to video search re-ranking. In the proposed approach, we take the text-based search results as the baseline for re-ranking. Then we exploit the conceptual as well as visual similarity to build virtual "hyperlinks" between video shots. By taking the video shots as the vertexes and the hyperlinks as the edges, we can construct a set of hierarchical graphs based on different semantic concepts. Upon these graphs, we apply a modified topic-sensitive PageRank algorithm to propagate the text-based relevance scores of video shots through the "hyperlinks" in each graph. The aggregated results of the propagated scores from the multiple graphs will be taken as the final ranking results of the search task.

This approach can be adapted to generic types of query as it is independent of query classes and requires no training data for query categorization. Also, it requires no involvement of human effort as the relevance of video shots to a given topic is propagated through the multiple graphs automatically. Furthermore, the fusion across textual, visual and semantic conceptual information is implemented in a graph-based iterative style, which combines the information from multimodalities in a natural and sound way. Evaluation of the proposed approach on TRECVID [40] dataset indicates that the graph-based propagation method of video search re-ranking significantly improves the performance of text-based search baseline.

On the other hand, as the baseline of the multimodal fusion, the text-based video search dominates the performance of the re-ranking approach. The existing IR (information retrieval) methods on plain text have been studied for many years. However, when applied to video search, these approaches are far from acceptable, although they are much mature and effective on text search tasks. The poor performance of text retrieval methods directly embedded in video search is due to the difference between the typical queries in video search and those in text search. For text search tasks, the queries are mostly semantic concepts (such as "web ontology" and "xml protocol"), the searching of which rely much upon their surface strings' relevance to the context of documents. Video search, however, is a task more content-and-visual based yet relatively less text-relevant. In video search tasks, queries are often "object-centric," inquiring for some visual objects, such as a person, an event and a scene. We name such objects as "targeted objects" in a query. Obviously, the query terms representing the "targeted objects" should be considered differently from those describing the background of the targeted objects.

Driven by this observation, we employ an approach to query analysis for improving the text-based search baseline as detailed in our work [26]. In this approach, we identify the "targeted objects" in a video search query and take special treatment to the query terms that represent the targeted objects. Specifically, we modify an object-centric BM25 algorithm, which emphasizes the contribution of specific query terms that represent the "targeted objects." In this way, we convert the text string query into an "*object query*." We name the proposed approach as "*object-sensitive query analysis*" for video search. In our proposed framework of video search system, a systematic yet minute query analysis process is placed before the text search stage to improve the search results. The improved text search results will be taken as the baseline of the multi-graph based multimodal fusion.

The rest of this paper is organized as follows. In Section 2, we present the object-sensitive query analysis approach to improve the text-based video search baseline. Section 3 details the proposed multi-graph based propagation method for video search re-ranking upon text-based search baseline results. Section 4 gives the experimental results of the proposed approaches on TRECVID dataset. In Section 5 we discuss the future work and conclude this paper.

2. TEXT-BASED SEARCH BASELINE

As aforementioned, text-based search result is an important baseline for video search. The graph-based propagation process is to update the states of the graphs in an iterative style, thus the performance of the propagation process relies much upon the initialization of the graphs, i.e. the search results from text-based search model.

To raise the bar of text-based search baseline, we employ an approach, namely "*object-sensitive query analysis*," [26] which significantly improves the text-based search results. Three steps, namely *N-gram query segmentation*, *name entity generalization*, and *object-sensitive query term re-weighting*, are applied to a query as a preprocessing stage. Specifically, in the step of object-sensitive query term re-weighting, we investigate four methods to identify the "targeted objects," namely *visual content-based semantic concept detection*, *part-of-speech (POS) identification*, *adverb refinement*, and *name entity reference highlight*.

The details of the object-sensitive query analysis approach to text-based search can be found in our work [26]. For the completion of the description of the proposed video search approach, we briefly review the query analysis approach in this section.

2.1 N-gram Query Segmentation

Before inputting the query topic string into the search engine, we first segment the query into term sequences based on N-gram method [3].

Given a query like "*find shots of one or more people reading a newspaper*" (a typical query in TRECVID search tasks), the key terms ("people," "read," and "newspaper" in this example) are retained after stemming (such as converting "reading" to "read") and stopwords (such as "a" and "of") removing. We apply the N-gram segmentation to the remained keywords. This particular example has three levels of N-gram (i.e., N is from 1 to 3). Therefore, seven query segments will be generalized as:

Unigram: $people^{(1)}, read^{(2)}, newspaper^{(3)};$

Bigram: $people\ read^{(4)}, read\ newspaper^{(5)}, people\ newspaper^{(6)};$

Trigram: $people\ read\ newspaper^{(7)}$.

These segments will be input to the search engine as different forms of the query, and the relevance scores of video shots retrieved by different query segments will be aggregated with different weights which can be set empirically. The higher gram a query segment has, the more relevant to the given query the corresponding video shots retrieved by this segment should be, and therefore the higher weight should be assigned. In this example, the video shots retrieved by "people read newspaper" will be given a higher aggregation weight than those retrieved by "people read."

2.2 Name Entity Generalization

Most queries for video search tasks contain the terms representing a name entity, such as a person, a place and a vehicle. In this paper, we employ a query expansion method for the refinement of queries with name entities. We name the method as "*name entity generalization*." In our approach, we classify the name entities into several predefined categories, and give each name entity a label of its corresponding category. The extraction of name entities and the application of the generalization method to query expansion are detailed as follows.

First, using an automatic name entity recognition tool [2], we identify the name entities occurring in both queries and text corpus associated with the video data. Then, a label of "*name entity category*" (such as "<person name>") is given to each identified name entity. For example, given a query "*find shots with one or more people leaving or entering a vehicle*," it will be tagged as: "*find shots with one or more people<person name> leaving or entering a vehicle<vehicle name>*." Similarly, we tag the name entities appearing in the text corpus of video data as well, e.g. "*Peter<person name> walks out of the car<vehicle name>*."

With this generalization method, name entities in both query and text corpus are tagged with the same set of category labels. Therefore, the relevant text segments which have no "direct" match to the original query will now be retrieved with these shared labels. As shown in the example above, the sentence which contains no query term before name entity generalization now can be retrieved by the labels which also occur in the expanded query.

2.3 Object-Sensitive Query Term Re-Weighting

2.3.1 Query Term Frequency

In general text search methods, all the query terms are treated equally, except that the term frequency in query (*qtf*) is taken into consideration, e.g. in BM25 [32] :

$$relevance = \sum_{T \epsilon Q} \omega \frac{(k_1+1)tf(k_3+1)qtf}{(K+tf)(k_3+qtf)} \quad (1)$$

where Q is a query consisting of term T; tf is the occurrence frequency of the term T within the text segment, qtf is the frequency of the term T within the topic from which Q was derived, and ω is the Robertson/Sparck Jones weight [34] of T in Q. K is calculated by:

$$K = k_1((1-b) + b * \frac{dl}{avdl}) \quad (2)$$

where *dl* and *avdl* denote the document length and the average document length, respectively. k_1, k_2 and b are empirically set parameters.

However, in the query of a video search task, *qtf* of all the terms is usually equal to "1," since there are rare terms occurring more than once in the query topic. Furthermore, merely using the query term frequency fails to consider the evidence of the semantic importance of different query terms. Therefore, to exploit the specific semantic characteristics of video queries and to better assess the importance of different query terms, we employ an object-sensitive query term re-weighting approach, which aims to distinguish the query terms representing the "targeted objects" from others representing the background of the targeted objects.

2.3.2 Identification of Targeted Object

To detect the "targeted objects" in a video search query, we define four identification methods which we name as: visual content-based semantic concept detection, POS (part-of-speech) identification, adverb refinement and name entity reference highlight, respectively.

Visual Content-Based Semantic Concept Detection

Content-based semantic concept detection is a widely used method for video annotation and retrieval. A semantic concept is an abstract description of the content of a video shot, for example, "person," "sports," etc. There are many public concept dictionaries, such as LSCOM [28] concept list which has become a general standard of concept detection and evaluation in the research community. It consists of more than 800 generic concepts, which represent the most important semantic concepts of video content. In our approach, we take LSCOM as the concept dictionary and compare each query term with the list. When there is a direct match between a query term and a concept of the list, the corresponding term is identified as a concept tag of the targeted video shots. Thus, it should be taken as the "targeted object" in the query.

Part-of-Speech Identification

In order to assess the syntactic characteristics of query terms, we construct POS (part-of-speech) tagging on the query with an automatic POS tagging tool [8]. Part-of-speech represents the syntactic property of a term, e.g. noun, verb, adjective, etc.

By labeling the query topic with POS tags, we can extract the terms with noun or noun phrase tags as the "targeted objects," as the noun and noun phrases often describe the centric objects that the query is inquiring for. For example, given a query "*find shots of one or more people reading a newspaper*," "people" and "newspaper" will be tagged as noun and extracted as the "targeted objects" in the query.

Adverb Refinement

Although extracted as "targeted objects," the noun and noun phrases at different positions of a sentence should be treated unequally due to their different importance. For example, the noun or noun phrases following an adverb with refinement meanings (such as "*with*" and "*at least*") represent the objects that must appear in the targeted video shots. We identify the adverbs with refinement meanings and take the noun or noun phrases following these adverbs as "targeted objects," e.g. the "*boats*" or "*ships*" in the query "*find shots of water with one or more boats or ships*."

Name Entity Reference Highlight

As aforementioned, name entities in the query can be identified with an automatic entity recognition tool. We observe that the different terms of a name entity do not always share the same occurrence rate. For example, in the reference of a publication, the author is more often referred by the last name rather than the first name. Based on such observation, we extract the underlying "targeted object" in name entities by identifying the part which is more often used as the reference of the name entity.

Take "George Bush" as an example. "*Bush*" occurs more often than "*George*" in the speech transcripts of broadcasted news when referring to "George Bush." And at most time, "*Bush*" refers to "George Bush" while "*George*" often refers to someone else. We calculate the frequency of different parts of a name entity from external data corpus, such as web search results, and select the most frequent part as the "targeted object" in the query.

2.3.3 Modified BM25 Algorithm

To emphasize the contribution of the terms representing "targeted objects" in the query, we define a modified qtf_{new} for BM25 equation (1):

$$qtf_{new} = \sum_i w_i * O_i(t) + qtf_{old} \tag{3}$$

$$O_i(t) = \begin{cases} 1 & if\ t\ is\ an\ targeted\ object; \\ 0 & otherwise. \end{cases} \tag{4}$$

where qtf_{old} represents the original query term frequency within the query topic as defined in (1). $O_i(t)$ represents the indicator function which predicts whether a term t represents a targeted object or not; w_i represents the weight assigned to the targeted object term detected by a specific identification method aforementioned (i = 1, 2, 3, 4). In special cases where a term is detected as the targeted object by more than one method, the scores from multiple methods will be aggregated and assigned to the term as a combined score. Specifically, in the case where the term is not detected as a targeted object by any method, the qtf_{new} will remain the same as the original query term frequency (qtf_{old}).

To combine the object-sensitive approach to query analysis with the text retrieval baseline in video search, we modify the original BM25 algorithm to an object-centric BM25 algorithm with the modification of *qtf* in equation (3) and (4):

$$relevance = \sum_{T \epsilon Q} \omega \frac{(k_1+1)tf(k_3+1)(\sum w*O(j)+qtf_{old})}{(K+tf)(k_3+\sum w*O(j)+qtf_{old})} \tag{5}$$

In the modified object-centric BM25 algorithm, not only the query term frequency is counted in, but also the object-based semantic importance of the query terms is taken into consideration.

With the object-sensitive query analysis approach, we enhance the performance of pure text-based methods employed in video search. Evaluation of the object-sensitive query analysis approach on TRECVID dataset shows significant improvement over the text search baseline [26].

3. VIDEO SEARCH RE-RANKING

Up to now, we have reviewed the object-sensitive query analysis approach to improving the text-based search baseline. The traditional multimodal fusion method in video search is typically a simple linear aggregation of search results from multimodalities, which does not exploit the underlying relationship between multimodalities. Furthermore, although the linear fusion method is easy to implement, much training data and human intelligence are required. Therefore, an optimized combining strategy is desired for a finer fusion of the information from multiple sources including textual, audio, semantic conceptual and visual features.

As aforementioned, we observed that there is an analogy between video shots and web pages: with the virtual "hyperlinks" indicating semantic relationships, video shots can construct a hierarchical structure similar to the hyperlinked web page structure. A straightforward intuition is that: by adopting a similar method to web page ranking utilizing hyperlinks, the video search problem can be addressed in a graph-based ranking fashion utilizing the hyperlinks of video shots as well.

Recently, the most widely used web page ranking algorithm is PageRank, which was proposed by Sergey Brin and Lawrence Page in 1998 [31]. In this paper, we propose a modified PageRank algorithm for video search re-ranking. To give a better explanation of the proposed algorithm, we first give a brief introduction of the PageRank algorithm and its modifications.

3.1 PageRank Algorithm

In [30], Nie et al. gave a detailed survey on the recent studies of PageRank algorithm, including the static PageRank, the dynamic PageRank, and the relevance-based intelligent surfer PageRank.

3.1.1 Static PageRank Algorithm

Brin and Page [31] proposed an alternative model of page importance, called the random surfer model. In that model, a surfer on a given page *i*, with probability *(1-d)* chooses to select uniformly one of its out-links *O(i)*, and with probability *d* jumps to a random page from the entire web *W*. The PageRank score for vertex (page) *i* is defined as the stationary probability of ending the random surfer at vertex *i*. One formulation of PageRank is given by:

$$PR(i) = (1-d) \sum_{j:j \to i} \frac{PR(j)}{O(j)} + d\frac{1}{N} \tag{6}$$

The static PageRank algorithm is a query-independent measure of the importance of web pages. It is only related to hyperlink structure of the entire web and has no bias to specific topics.

3.1.2 Dynamic PageRank Algorithm

In Haveliwala's Topic-Sensitive PageRank (TSPR) [15], a set of topics consisting of the top level categories of the ODP [15] are selected, with τ_i as the set of URLs within topic c_j . Multiple PageRank calculations are performed on each topic, respectively. When computing the PageRank vector for topic c_j , the random surfer will jump to a page in τ_i at random rather than just to any page in the whole web. This has the effect of biasing the PageRank to that topic. Thus, page *k*'s score on topic c_j can be defined as:

$$TSPR_j(k) = (1-d) \sum_{i:i \to k} \frac{TSPR_j(i)}{O(i)} + d\frac{1}{N} \tag{7}$$

To rank results for a particular query *q*, let *r(q, c_j)* be *q*'s relevance to topic c_j . For web page *k*, the query sensitive importance score is given by:

$$S_q(k) = \sum_j TSPR_j(k) * r(q, c_j) \tag{8}$$

The relevance results of web pages to a given query are ranked according to this composite score.

3.1.3 The Intelligent Surfer

Richardson and Domingos [33] proposed an intelligent surfer PageRank algorithm (ISPR), in which the surfer is prescient, selecting links (or jumps) based on the relevance of the target to the query of interest. In such a query-specific version of PageRank, the surfer still has two choices: follow a link, with probability *(1 – d)*, or jump with probability *d*. However, instead of selecting among the possible destinations equally, the surfer chooses the target using a probability distribution generated from the relevance of the target to the surfer's query. Thus, for a specific query *q*, page *j*'s query-dependent score can be calculated by:

$$IS_q(j) = d\frac{r(q,j)}{\sum_{k \epsilon W} r(q,k)} + (1-d) \sum_{i:i \to j} IS_q(i) \frac{r(q,j)}{\sum_{l:i->l} r(q,l)} \tag{9}$$

3.2 Multi-Graph Construction

In this paper, we will formulate the video search problem in a graph-based fashion, by exploiting the analogy between video shots and web pages in a "*PageRank-like*" way. Based on this viewpoint, we will construct hyperlinked graphs of video shots similar to those of web pages. Then we will apply a modified topic-sensitive PageRank algorithm to propagate the relevance scores of video shots through these graphs. The video shots will then be re-ranked according to the aggregation scores of the multi-graph based propagation.

In the following paragraphs, we will demonstrate the process of video search in the "PageRank-like" way by constructing the hyperlinked graphs of video shots.

3.2.1 Text-Based Search Model

Text-based search model is the baseline of most multimodal fusion methods currently studied. We also take the text-based search results as the baseline of our multi-graph re-ranking model.

A more formal definition of text retrieval in video search problem is: given a query in text, we are to estimate the relevance R(x) of each video shot x in the search set X ($x \in$X) to the query, and order them by their relevance scores. The relevance of a shot is given by the relevance score between the associated text of the shot and the given text query.

With the text-based search model presented in Section 2, each video shot is assigned with a relevance score on the given text query. The higher relevance score, the higher likelihood that the shot is related to the given query.

Given the retrieved video shots and their relevance scores, we could treat the video shots in a similar way to the retrieved web pages in a web search task. We take the video shots as vertexes, and construct a vertex-weighted graph with these video shots. The text-relevance score of each shot is considered as the weight of each vertex, similar to the relevance score of each web page to the given topic in a web search task. The video shots that are irrelevant to the query (identified by text-based search model) have a default relevance score equal to zero. An exemplary graph of a set of video shots is shown in Fig. 2.

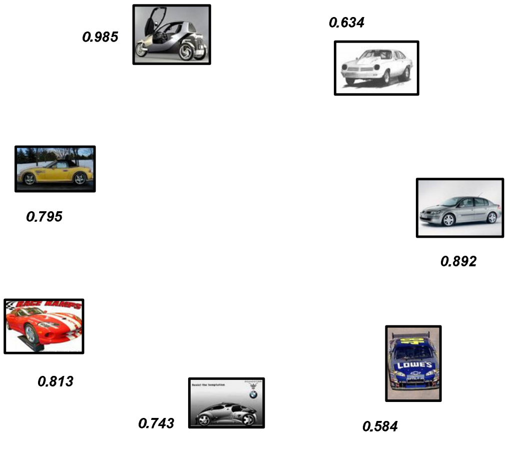

Fig. 2. A graph with video shots as vertexes.

3.2.2 Concept Detection Model

Semantic concept detection is a widely studied topic in multimedia research areas [4][5][29][39][43]. A concept detection model predicts the likelihood of a video shot being related to a given concept, and classifies the video shots into positive category (relevant) and negative category (irrelevant) on a given concept.

In our approach, we employ the concept detection model to assess the virtual semantic relations between video shots. We use several models to implement concept detection, such as SVM (Support Vector Machines) [20][42], manifold ranking [45] and transductive graph [10]. The details of these models for concept detection can be found in [16]. Briefly speaking, these models detect the relevance of each video shot to a specific concept, and rank the video shots according to their "confidence scores" of being relevant to the concept.

With the concept detection model, we can get a set of relevant video shots to each concept respectively. The set of relevant video shots to a specific concept are not independent of each other, but share some semantic relationship. This relationship is similar to the case of web pages. A pair of web pages which have a "hyperlink" between each other share some semantic relationship, which is indicated by the anchor texts of the hyperlink. Similarly, the concept to which a set of video shots are related indicates the semantic meanings of the contents of these video shots. Therefore, the semantic meaning which is shared by a pair of video shots can be taken as the "hyperlink" between each other as well, with the corresponding concept as the "anchor text."

Given a query, we can select a set of concepts that are highly relevant to the query from a concept dictionary. The relevant concepts to a given query can be retrieved through typical text processing methods, such as surface-string similarity computation [27][41], context similarity comparison [35], ontology and dictionary matching (such as using WordNet [11]). For each concept mapped to the query, we can obtain from the concept detection model a set of video shots which are relevant to the concept. Then we build a virtual "hyperlink" between each pair of these video shots indicating that the two shots have a semantic concept similarity.

Thus, for the set of concepts mapped to the given query, there will be a set of graphs constructed based on individual concepts. Each graph consists of all the video shots that are relevant to the corresponding concept. Fig. 3 illustrates an exemplary graph constructed on a specific concept "car." The vertexes of the graph are video shots that are relevant to the concept "car." Each vertex contains a text-relevance score generated from the text-based search model, as well as a confidence score of being relevant to the concept "car" generated from the concept detection model. This graph indicates that there is a semantic concept similarity between each pair of the "hyperlinked" video shots, and the similarity refers to the concept "car."

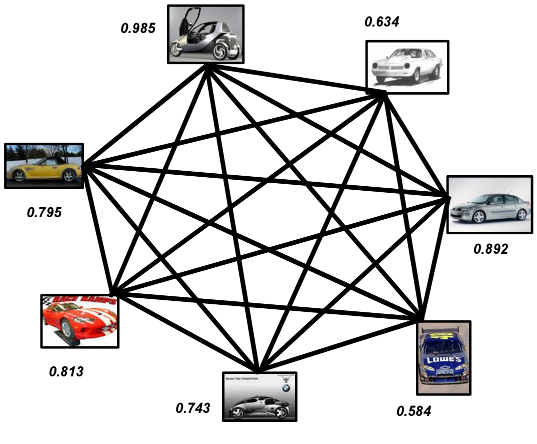

Fig. 3. A graph based on a specific concept "car."

3.2.3 Visual Similarity Model

The assumption we adopt in the graph construction procedure is that, if two video shots are predicted as positive instances by concept detection model, they probably share a semantic conceptual similarity between each other. However, due to the limited performance of concept detection methods, two shots which are both predicted as relevant to a concept may have no similarity actually. Therefore, to better reinforce the relationship between video shots by tightening the constraint of "hyperlinks" generated from wrong prediction, we have to exploit other information besides semantic concept similarity into the graph construction.

A widely used similarity measure of video shots is the content-based visual similarity, which can be obtained from low-level features of video shots. In our approach, we employ a visual similarity comparison model of these low-level features to refine the hyperlinks in the "PageRank-like" graphs of video shots.

The comparison model of visual similarity is implemented as follows: we build a vector for each video shot with low-level visual features (where visual features on color moment are mainly used) as the vector elements. Then for each pair of video shots, we compare the distance of the corresponding pair of vectors ($Distance(X_i, X_j)$), and take it as the measure of visual similarity of video shots. One form of the distance equation is aggregating the divergence of feature values on each dimension:

$$Distance(X_i, X_j) = \sum_d |x_{id} - x_{jd}| \tag{10}$$

where x_{id} is the value of the d-th element of the feature vector of video shot i, i.e. the d-th low-level feature of shot i.

Then we give a threshold of distance to filter the video shot pairs which have low visual similarity. Only those with a distance smaller than the threshold are taken as similar pairs. And the hyperlink between a pair of video shots which share a distance larger than the threshold will be taken as pseudo-pairs and then pruned from the "PageRank-like" graph.

Fig. 4 gives an illustration of the graph pruned from the aforementioned exemplary graph constructed based on the concept "car" (Fig. 3). After pruning, the complete graph constructed by the concept detection model is now modified to an incomplete graph, with only the hyperlinks connecting highly relevant pairs of video shots retained.

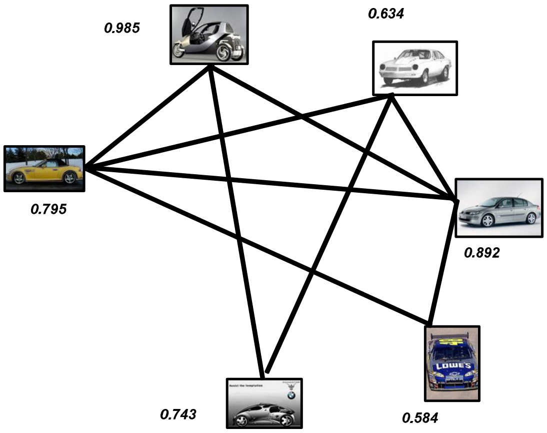

Fig. 4. A graph pruned based on visual similarity.

3.2.4 Edge Direction Assignment

In the web space, a pair of web pages which are connected by a hyperlink do not always have the same importance, especially on a specific topic. The kernel assumption in PageRank algorithm is that, the web page "in-linked" by a hyperlink has a higher importance than the web page "out-linked" by the hyperlink, as a more important web page is theoretically cited more frequently than other less important ones. Similarly, although sharing a mutual relationship of conceptual and visual similarity, two video shots connected by a "hyperlink" in the "PageRank-like" graph do not always have the same importance in the video shot space as well.

"Random walk" is another assumption in PageRank algorithm. It is assumed that Internet surfers will "random walk" to a web page following the hyperlinks within the current web page, or randomly "jump" to a web page out of the linked set. Although the walking or jumping behavior is random, the web pages which are in-linked by more hyperlinks will have a larger probability to be visited than others which have less in-links.

This "random walk" idea can be ported into video search as well. We assume the video shots retrieved by search models as a set of web pages in a web space. Therefore, when a user "surfs" among the video shots for a given query, he will "random walk" to another video shot which is in-linked by this video shot, or jump to a video shot which has no hyperlinks with the current shot. However, the probability of "walking" to an in-linked video shot is much larger, as a video shot that is more relevant to the query (in-linked by the current video shot) has a larger chance to be visited rather than other unlinked video shots. The reason is that the user has a query in mind, and is searching for relevant video shots. Thus, when he finds a relevant video shot to the query, he will prefer to follow the out-link of this video shot to a more relevant shot, in order to reach the targeted video shots.

As a concept related to the given query is a bridge between the video shots and the query, the video shot which contains a higher confidence score of concept detection on this specific concept is more relevant to the query than a shot that has a lower confidence score. Therefore, we can assign a direction between each pair of video shots by comparing the confidence scores of these video shots from concept detection models. The direction is assigned as: the hyperlink will be "out-linked" from the video shot with lower confidence score to the one with higher confidence score, so that a surfer following the out-link of a video shot will reach to a more relevant shot.

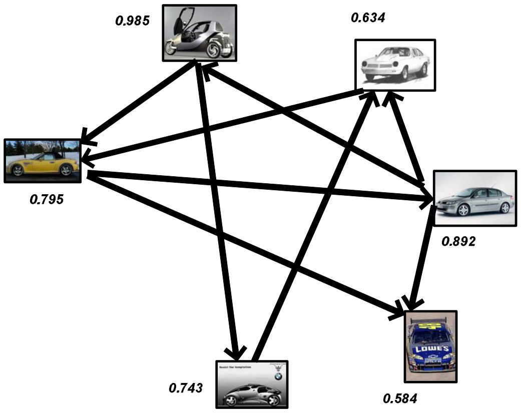

Fig. 5. A graph re-constructed with directed hyperlinks.

Fig. 5 shows an illustration of a directed graph. For each edge in the pruned graph in Fig 4, a direction is assigned from the shot with lower concept confidence score to that with higher score, i.e., the vertex that is more relevant to the given topic is "in-linked" by the hyperlink and that the one less relevant is "out-linked" by the hyperlink.

3.3 Video-PageRank Algorithm

Up to now, we have exploited the underlying conceptual and visual similarity relationships between video shots, and simulate the video search problem in a "PageRank fashion." In summary, we construct a uni-graph based on a specific concept in the following procedure: vertex weighting by text-based search model, hyperlink construction by concept detection model, graph pruning by visual similarity comparison model, and hyperlink direction assignment with confidence score from concept detection model.

Moreover, given a set of concepts related to a given query, we can construct a set of graphs based on each individual concept. Upon the multiple graphs, we apply a modified "intelligent surfer" PageRank (ISPR) algorithm for video search and propose a graph-based propagation approach to re-ranking the text-based search results. In this paper, we name the proposed "Intelligent Surfer" PageRank algorithm for Video Search as ISPR-VS algorithm.

The ISPR-VS algorithm is explained as follows. We assume that a surfer (similar to a surfer in the web space) is browsing among a graph of video shots and searching for relevant video shots to a given query *q*. At a specific video shot *j*, the surfer will choose to select one of the out-links of the current shot uniformly, or jump to a video shot in the entire video corpus randomly. For the next step of browsing, the surfer has two choices: follow a link, with probability *(1 − d)*, or jump, with probability *d*. However, the surfer in a video search task is prescient rather than random walking, as the text-relevance score of each video shot to the query is provided as priori-knowledge. Therefore, the surfer will select the links (or jump) based on his/her interest of query. Instead of selecting among the possible destinations uniformly, the surfer chooses using a probability distribution ($\frac{ASR(q,j)}{\sum_{k \epsilon G} ASR(q,k)}$), where *ASR(q,j)* refers to the ASR-based text relevance score of the targeted video shot to the surfer's query. ASR refers to automatic speech recognition, which is widely employed to generate text corpus associated with video data from embedded audio speech.

The ISPR-VS score calculated from the graph constructed on a specific concept *c* is given by:

$$IS_{q,c}(j) = \begin{cases} d\frac{ASR(q,j)}{\sum_{k \epsilon G(c)} ASR(q,k)} + (1-d)\sum_{i:i \to j\,(c)} IS_{q,c}(i)\frac{ASR(q,j)}{\sum_{l:i \to l} ASR(q,l)} \\ d\frac{ASR(q,j)}{\sum_{k \epsilon G(c)} ASR(q,k)}, if\ shot\ j\ doesn't\ map\ to\ the\ concept \end{cases} \quad (11)$$

where *ASR(q,j)* represents the ASR-relevance score of shot *j* to the given query *q*, generated from the text-based search model. *G(c)* represents all the video shots in the graph generated on concept *c*. *d* is a parameter similar to that in the static PageRank algorithm, which can be set empirically. *l* represents the shots that out-link to the shot *j* in the graph constructed based on concept *c*, i.e., *l* represents the shots that have lower concept confidence score than shot *j* on the concept *c*. For the shot that has no relevance to the concept *c*, an initial text-relevance-based score is given to the shot ($d\frac{ASR(q,j)}{\sum_{k \epsilon G(c)} ASR(q,k)}$).

Thus, for a specific query *q*, video shot *j*'s query-dependent score within the graph based on a specific concept *c* can be calculated as $IS_{q,c}(j)$. This re-ranked relevance score will be propagated on each video shot iteratively until convergence, as the ISPR-VS algorithm is recursive.

Based on the propagation, we further define an aggregation algorithm upon multiple graphs. The aggregated score of multi-graph propagation is given by:

$$IS_q(j) = \sum_c IS_{q,c}(j) \quad (12)$$

where $IS_{q,c}(j)$ represents the relevance score of video shot *j* to the query within the graph based on concept *c*. $IS_q(j)$ denotes a linear combination of all the $IS_{q,c}(j)$ scores on the set of query-related concepts. With this combination, the aggregated relevance scores of video shots will be taken as the final re-ranking results.

4. EXPERIMENTS

To verify the effectiveness of the proposed approaches, we apply them to the data set of TRECVID 2006 [40], which consists of 259 videos with 79,846 shots in a size of 87.7G. 24 queries are provided for the search task.

In the experiments that evaluate text-based search model, we take the original BM25 algorithm as the baseline, and incrementally add the proposed object-sensitive query analysis methods to the baseline algorithm. Table 1 shows the experimental results of average MAP on all search tasks.

From Table 1 we could see that the object-sensitive query term re-weighting method improves the performance of text-based search baseline significantly. The N-gram query segmentation and the name entity generalization methods, although have relatively smaller contribution, also improve the performance. More detailed experimental results can be found in our work [26].

Table 1. Performance of object-sensitive query analysis.

Approach	Average MAP
Pure text search baseline (BM25)	0.0328
+ N-gram query segmentation	0.0345
+ Name entity generalization	0.0367
+ Object-sensitive query term re-weighting (modified BM25)	0.0424

Then we take the text-based search results as the baseline for re-ranking and implement the graph-based propagation. For the mapping of query and concept, we use the concept list of the LSCOM [28] as the concept dictionary. When there is a direct match between a query term and a concept of the list, or when the query belongs to a general category represented by a concept (e.g. "person," "sports"), the corresponding concept is identified as relevant to the query and a graph of video shots is constructed based on this concept.

Table 2. Performance of graph-based propagation approach.

Approach	Average MAP	Gain
Pure text search baseline (BM25)	0.0328	-
Our Text search baseline (modified BM25)	0.0424	29.3%
+ Multi-graph based re-ranking	0.0651	53.6%

As the performance of propagation relies much upon the results from the concept detection model, to get a stable performance of propagation as well as to speed up the efficiency of iteration, we only choose the concepts which have high accuracy in concept detection. Therefore, we evaluate the concept detection results with the ground truth provided by TRECVID2006 high-level feature extraction task, and filter the concepts that get low precision. The experimental results of the multi-graph based propagation approach are shown in Table 2.

For the PageRank-like algorithm, we set the parameter *d* equal to 0.85, which is taken as the empirical setting in the classic PageRank algorithm. Experimental results show that the multi-graph based re-ranking approach gets an average MAP of 0.0651, which significantly outperforms the text-based search baseline (0.0424) by 53.6%.

Fig. 6 shows the comparison of the proposed re-ranking approach and the text-based search baseline on each query of TRECVID2006 dataset. The experimental results indicate that with graph-based re-ranking, the performance of video search is improved upon the text-based search baseline on most queries, although suffers slight loss on several queries, such as query "177" and "194."

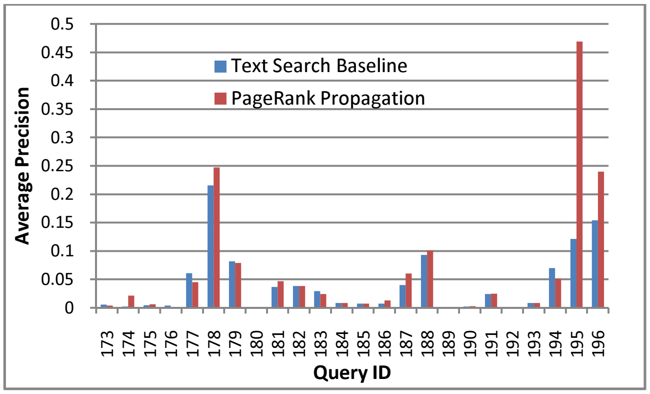

Fig. 6. Performance of graph-based propagation.

5. CONCLUSIONS & FUTURE WORK

In this paper we addressed the problem of multimodal re-ranking in video search. We proposed a graph-based propagation method to re-rank the relevance score of video shots. In the proposed approach, we investigated the video search problem in a "PageRank fashion," by taking the video shots as web pages and the conceptual as well as visual similarity as hyperlinks. Then we proposed a modified topic-sensitive PageRank algorithm to propagate the relevance score of video shots through these hyperlinks. We also employed an object-sensitive approach to query analysis with a modified BM25 algorithm to improve the performance of text-based search baseline. Through experiments, we verified the effectiveness of the proposed graph-based re-ranking approach combining with the object-sensitive approach on the dataset of TRECVID2006. Experimental results indicated that the proposed video search approaches significantly improve the performance of pure text-based video search results.

For future work, we will focus on the study of applying other graph-based propagation methods to video search, such as HITS algorithm, to better implement video search re-ranking task. We are also to tackle the problem of identifying the targeted objects from multimedia data rather than only plain text, to improve the object-sensitive query analysis method.

6. REFERENCES

[1] Amir, A. et al. IBM Research TRECVID-2005 video retrieval system. In TRECVID Workshop, Washington DC, 2005.

[2] Alias-i. Lingpipe named entity tagger. In http://www.alias-i.com/lingpipe/.

[3] Brown, Peter F., deSouza, Peter V., Mercer, Robert L., Della Pietra, Vincent J., Lai, Jenifer C. Class-Based n-gram Models of Natural Language. Computational Linguistics, 1992

[4] Chang, Shih-Fu et al. Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction. In TRECVID Workshop, 2006.

[5] Chang, Shih-Fu et al. Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction. In TRECVID Workshop, 2005.

[6] Chang, T.-C et al. TRECVID 2004 Search and Feature Extraction Task by NUS PRIS. In TRECVID Workshop, Washington DC, 2004.

[7] Campbell, M., Haubold, A., Ebadollahi, S., Naphade, Milind R., Natsev, A., Smith, John R., Tesic, J., Xie, L. IBM Research TRECVID-2006 Video Retrieval System. In TRECVID Workshop, 2006.

[8] CLAWS part-of-speech tagger for English. http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/

[9] Donald, K. M. and Smeaton, A. F. A Comparison of Score, Rank and Probability-Based Fusion Methods for Video Shot Retrieval. International Conference on Image and Video Retrieval (CIVR), 2005.

[10] El-Yaniv, R., and Gerzon, L. Effective Transductive Learning via PAC-Bayesian Model Selection, Technical Report CS-2004-05, Technion-Israel Institute of Technology, 2004.

[11] Fellbaum, Christiane. WordNet: an Electronic Lexical Database, MIT Press, 1998.

[12] Hsu, W. H. and Chang, S.-F. Topic Tracking Across Broadcast News Videos with Visual Duplicates and Semantic Concepts. In International Conference on Image Processing (ICIP), Atlanta, GA, USA, 2006.

[13] Hauptmann, A. G. and Christel, M. G. Successful Approaches in the TREC Video Retrieval Evaluations. ACM Multimedia 2004.

[14] Hauptmann, A.G., Chen, M.-Y., Christel, M., Lin, W.-H., Yan, R., Yang, J. Multi-Lingual Broadcast News Retrieval. TRECVID2006.

[15] Haveliwala, Taher H. Topic Sensitive PageRank, International World Wide Web Conference (WWW), 2002.

[16] Hua, X.S., Mei, T., Lai, W., Wang, M., Tang, J., Qi, G.J., Li, L., Gu, Z. Microsoft Research Asia TRECVID 2006 High-Level Feature Extraction and Rushes Exploitation. TRECVID 2006.

[17] Hauptmann, Alexander G., Lin, W.H., Yan, R., Yang, J. and Chen, M.Y. Extreme Video Retrieval: Joint Maximization of Human and Computer Performance, ACM Multimedia 2006.

[18] Hsu. W. H. and Chang. S.-F. Visual Cue Cluster Construction via Information Bottleneck Principle and Kernel Density Estimation. In the International Conference on Image and Video Retrieval (CIVR), Singapore, 2005.

[19] Hsu. Winston H., Kennedy, Lyndon S., Chang, Shih-Fu. Video Search Reranking via Information Bottleneck Principle. ACM Multimedia 2006.

[20] Herbrich, R., Graepel, T., and Obermayer, K. Support Vector Learning for Ordinal Regression. In Proc. of the 9th International Conference on Artificial Neural Networks, 1999.

[21] Iyengar, et al. Joint Visual-Text Modeling for Automatic Retrieval of Multimedia Documents, ACM Multimedia 2006.

[22] Joachims, Thorsten. 2004. SVMlight -- Support Vector Machine. http://svmlight.joachims.org/.

[23] Kennedy, L., Natsev, P., and Chang, S.-F. Automatic Discovery of Query Class Dependent Models for Multimodal Search. In ACM Multimedia, Singapore, 2005.

[24] Kurland, Oren and Lee, Lillian. PageRank Without Hyperlinks: Structural Re-Ranking Using Links Induced by Language Models, Proceedings of the 28th ACM SIGIR conference on Research and Development in Information Retrieval, 2005.

[25] Kurland, Oren and Lee, Lillian. Respect My Authority! HITS Without Hyperlinks, Utilizing Cluster-Based Language Models.SIGIR'06

[26] Liu, J., Hua, X.S., Li, S. Object-Sensitive Query Analysis for Video Search. Multimedia Signal Processing, 2007.

[27] Lin, Dekang. Automatic Retrieval and Clustering of Similar Words. COLING-ACL'98.

[28] LSCOM Lexicon Definitions and Annotations (Version 1.0). DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia. http://www.ee.columbia.edu/ln/dvmm/lscom/

[29] Natsev, A., Naphade, M. R., and Tesic, J. Learning the Semantics of Multimedia Queries and Concepts from a Small Number of Examples. In ACM Multimedia, Singapore, 2005.

[30] Nie, L., Davison, Brian D., Qi, X. Topical Link Analysis for Web Search. SIGIR'06

[31] Page, L., Brin, S., Motwani, R., and Winograd, T. The Pagerank Citation Ranking: Bringing Order to the web. Technical report, Stanford University, Stanford, CA, 1998.

[32] Robertson, S. E. Overview of the Okapi Projects, Journal of Documentation, Vol. 53, No. 1, pp. 3-7, 1997.

[33] Richardson, M. and Domingos, P. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. In Advances in Neural Information Processing Systems 14. MIT Press, 2002.

[34] Robertson, S. E., and Sparck Jones, K. Relevance Weighting of Search Terms. Journal of the American Society of Information Science, 1976.

[35] Salton, G., Wong, A., and Yang, C. S. A Vector Space Model for Automatic Indexing, Communications of the ACM, 1975

[36] Shipman, F., Girgensohn, A., and Wilcox, L. Hyper-Hitchcock: Towards the Easy Authoring of Interactive Video. Human-Computer Interaction INTERACT '03, IOS Press, pp. 33-40, September 1, 2003.

[37] Shipman, F., Girgensohn, A., and Wilcox, L. Combining Spatial and Navigational Structure in the Hyper-Hitchcock Hypervideo Editor. Proceedings of Hypertext '03, pp. 124-125, August 26, 2003.

[38] Shipman, F., Girgensohn, A., and Wilcox, L. Creating Navigable Multi-Level Video Summaries IEEE International Conference on Multimedia and Expo, v. II, pp. 753-756, 2003.

[39] Snoek, Cees G.M., Worring, M., van Gemert, Jan C., Geusebroek, J.M., and Smeulders, Arnold W.M. The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. ACM Multimedia 2006.

[40] TRECVID. TREC Video Retrieval Evaluation. In http://www-nlpir.nist.gov/projects/trecvid/.

[41] Ukkonen, E. Algorithms for Approximate String Matching. Information and Control, 1985.

[42] Vapnik, V. N. Statistical Learning Theory, John Wiley & Sons, 1995.

[43] Wu, Y., Tseng, Belle. L., Smith, John R. Ontology-based Multi-Classification Learning for Video Concept Detection. International Conference on Multimedia & Expo (ICME), 2004.

[44] Wang, C., Jing, F., Zhang, L., Zhang, H.J. Image Annotation Refinement using Random Walk with Restarts, ACM Multimedia 2006.

[45] Zhou, D., Bousquet, O., Lal, T., Weston, J., and Scholkopf, B. Learning with Local and Global Consistency, in Proc. Advances in Neural Information Processing System, 2004.