# Identifying Authoritative Sources of Multimedia Content *

Lyndon Kennedy
Yahoo! Labs
4301 Great America Parkway
Santa Clara, CA 95054
lyndonk@yahoo-inc.com

Malcolm Slaney
Yahoo! Research
4301 Great America Parkway
Santa Clara, CA 95054
malcolm@ieee.org

## ABSTRACT

We present a framework for identifying authoritative sources (such as web sites or individual users) that are likely to produce high-quality or interesting images. We construct a directed graph across sources based on the propensity of one source to "cite" the content from another. A graph-centrality measure scores the authority for each source, which could then be applied for retrieval purposes. We apply this method to web image retrieval, where web sites are the sources, and citations are found via copy detection; and on a photo sharing site, where individuals are the sources and citations are users' favorites. We are able to identify primary or influential sources of media while avoiding the computational cost of other approaches.

## Categories and Subject Descriptors

H.3.1 [**Information Search and Retrieval**]: Content Analysis and Indexing

## General Terms

Algorithms, Human Factors

## 1. INTRODUCTION

Multimedia retrieval is a challenging problem. Millions of images are added to the web every day and users increasingly require mechanisms for navigating these massive collections. The challenge is that the actual information contained in images and videos (matrices of pixels and streams of audio) does little to reveal the semantic meaning of the media.

There are four popular signals for ranking objects: image features, the text around the object, the (web) links, and popularity. In this paper we propose another signal based on citations.

1) Pioneering approaches to visual search relied upon content cues extracted from the image itself: typically distributions of low-level features, such as colors, textures, and edges in the images. These signals are difficult to exploit due to the "semantic gap" between content and high-level meaning.

2) Currently web-scale multimedia search uses text around the object. This works because a web page containing an image often contains words describing the content. But this information doesn't necessarily tell us which image of the Golden Gate Bridge, for example, is the best one.

3) People often credit PageRank [7] for making it possible for users to navigate the modern web. The key insight is that a hyperlink is essentially an editorial judgement by a real person that the linked-to document is somehow important and PageRank provides a mechanism for aggregating these recommendations from millions of web authors.

4) Objects can also be ranked based on popularity. This is important since most multimedia objects do not come with links. Flickr, for example, uses a (secret) combination of view counts and lists of personal favorites to create a measure of interestingness. Then one can sort images that match a keyword by interestingness or popularity.

In this work, we propose a new cue based on how multimedia documents are *produced*. We focus on the *source* of the document, meaning the specific web site, user, or IP address that is providing the piece of media. We aim to characterize how *authoritative* this source is and how likely it is to provide media that are frequently relevant, interesting, or otherwise *reliably* relevant.

The primary contribution of this work is a new cue for measuring image citation and a framework for leveraging this cue to determine the importance of the *source* of an image. We propose that authoritative sources can be found via an adaptation of traditional analyses of citation networks. If we can construct a graph of sources and their relative propensities to "cite" each other, then we can apply ranking techniques, such as PageRank [7], to determine the authority of each source. We evaluate the proposed framework by examining the propensity of web sites to copy images from each other. We define this copying behavior as a form of citation and find that we are able to identify sites that are often primary sources for iconic images. We also show that users marking photos as favorites on social media sites is a signal which can be used to identify authoritative users.

## 2. RELATED WORK

An important piece of related work is the well-known PageRank algorithm [7]. In its initial application, PageRank computes the authority of web pages by interpreting hyperlinks as a sort of citation. The web pages are taken as nodes and the hyperlinks between them as directed edges and the

*Area chair: Bernard Merialdo


*MM'11,* November 28–December 1, 2011, Scottsdale, Arizona, USA.

resulting ranking of pages is roughly the stationary probability of a random walk over that graph. This approach is not limited to only web pages and hyperlinks. Several previous works have shown applications of this technique for multimedia retrieval. In these works, the nodes are multimedia documents in a search result, such as video clips [1], web images [3], or video stills [5] and the edges between the nodes are weighted by the visual similarity between them. Applying a random walk technique, like PageRank, over these types of graphs shows improvements in the overall relative ranking of documents.

Some prior work has looked at creating the graph structure not at the level of the document, itself, but at the level of the *sources* of the documents. These "SiteRank" solutions are applied to web documents, where web pages are treated as documents and the *sites* (typically the domain names from which the pages come) as the sources. The relative strength of links between any two sources is typically weighted by the number of hypertext documents in one source with hyperlinks to documents from the target source. The resulting ranking of sources can be applied to making crawling mechanisms more efficient [2], enabling persistent search systems [8], or browsing peer-to-peer networks [10].

Our proposed approach is different from previous work in that we propose to characterize multimedia documents not by their individual content, but by the quality of the sources from which they are drawn. We believe this is the first work to apply random-walk techniques to multimedia sources rather than the documents themselves.

# 3. PROPOSED FRAMEWORK

In our proposed system, we construct a citation network between a set of sources. "Sources" are persons, publications, or any other entities that produce, disseminate, or consume multimedia content and a "citation" occurs when one source references a piece of media from another source. We reason that some sources are more likely to be producers of high-quality, relevant, or interesting multimedia documents than others and we hypothesize that this quality will be reflected by the centrality of the source in this resulting source citation network. Intuitively, sources that are frequently cited are likely to continually produce high-quality media.

We construct the citation network such that each node represents a source and edges are placed between nodes and weighted based on the propensity of one source to cite another source. Edges are directed towards the source being cited and the weights of all outgoing edges for each node are normalized to sum to one. We then apply the PageRank algorithm across this network, which is essentially a random walk over the citation graph with random restarts. The stationary probability distribution of this random walk over nodes is then interpreted as a ranking of each of the sources according to their relative authority.

If we conduct the process as described above over a heterogeneous collection of sources, then we will ultimately arrive at a query-independent ranking of sources. However, some sources may be more specialized than others, meaning that this general ranking of authority might not be applicable for all types of queries. We then propose to further improve the approach by enabling ourselves to find more-specific subgraphs, where the sources are likely to be constrained to certain classes of topics. If we apply PageRank to these subgraphs, then we will expect the resulting rankings of nodes to be different from the general case and unique to the specific topics of the sources that we have sub-selected. We call these analyses over topic-sensitive subgraphs query-class-dependent rankings, as they can be utilized separately based on the type of query being addressed.

# 4. APPLICATION TO IMAGE SEARCH

A key insight of the original PageRank work is that hyperlinks between websites are functionally equivalent to citations. That is, a link from one website to another is essentially a vote of confidence conferred upon the target page. When website authors wish to mention or utilize an image, however, they do not typically just create a hyperlink to the original. Instead, they will make a copy of the image and re-post it. We observe that these copied images are effectively citations of the original image and that the most frequently copied images on the web are often more likely to be relevant to the query and subjectively of higher quality or more iconic. In this application, we go a level higher to the *sources* that are actually providing the images. In this case, the sources are individual web sites and we aggregate individual cases of image copying across various websites into an overall propensity for citation between these sources.

## 4.1 Overall Approach

Our approach to tracing citation patterns across sources relies heavily on the detection of near-duplicate images across the entire web. With these image near-duplicate detection results, we would like to aggregate inter-source citations and construct a source citation graph across which to conduct PageRank. This is, unfortunately, infeasible across all of the billions of images available on the web. Instead, we propose to populate the lists of possible sources by issuing queries against a commercial web image search engine and to detect near-duplicates within the resulting images returned. Specifically, we limit near-duplicate detections to be conducted solely within the images returned for each query, which greatly reduces the computational complexity with a very small expense of decreased recall in the discovery of near-duplicate images: multiple copies of the same image are far more likely to be returned within the same query rather than across disparate queries.

## 4.2 Citation Detection

A citation occurs when an image is copied and reposted elsewhere. To detect this, we have to detect whether two images are copies of each other and if they are, then resolve which one is citing which.

Copy detection has seen a great deal of interest and a number of workable proposed solutions. In this work, we opt for a simple approach, which matches scale-invariant feature transform (SIFT) [6] descriptors across candidate image pairs. A version of this method is included in a freely-available SIFT toolkit [9].

Given two instances of an image, both of which are detected as copies of each other, we are still left with the question of which image was copied from which, or more pointedly, which source is the originator of the image and which source is merely citing that original image. We have solved this with a very simple approach: we extract the 'Last-Modified' date from the HTTP headers for the images and project that the older file is being cited by the newer file. This, of course, may not actually reflect the reality of

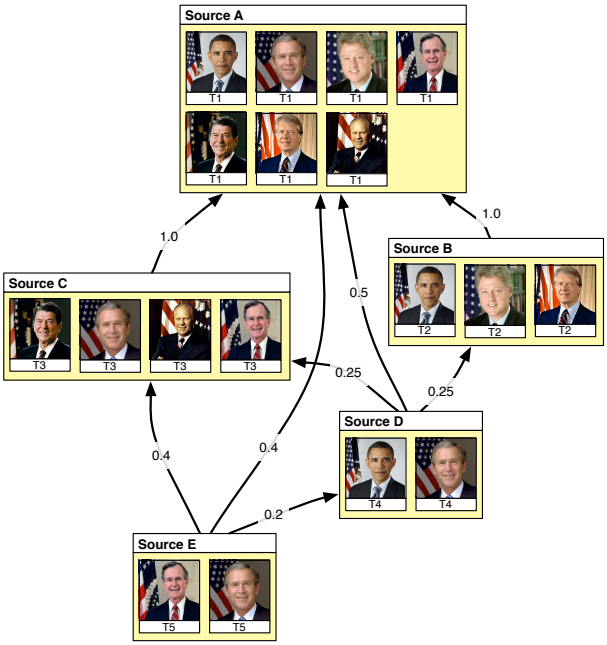


**Figure 1: A source citation network is extracted from repeated images across multiple sources. The edges from a source are weighted according to how many images are being reused between the two sources. The out-links from one source are normalized to sum to one.**

how the images were discovered and repurposed, but it is a reasonable approximation.

## 4.3 Graph Construction

Once we have pairwise citations detected between individual images, we need to move on to constructing a graph between sources. The graph consists of sources (nodes) and citation propensities (weighted, directed edges). For each pair of near-duplicate images that we have detected, we draw an edge between the sources (domains) of the images, directed from the citing source to the original source. If multiple such pairs are discovered between two sources, the weight of the edge between the two sources is weighted proportionally. Finally, the weights of the edges are normalized such that the outlink weights from each node sum to unity.

To illustrate this citation network construction process, Figure 1 shows a hypothetical source graph across sources of images for recent United States presidents. There are seven images in total dispersed across five different sources. In each source, the image is also tagged with a timestamp. (For simplicity, the times are simply T1-T5 and all images within a source have identical timestamps.) The weight of the edges between any two sources is proportional to the amount of images that they have in common between them.

The queries used to seed the graph construction dictate the types of authorities that we will find. We propose to leverage this to arrive at different authority scores for different classes of queries. For example, if we choose to seed the model with only queries for professional athletes, then the resulting citation graph structure will be highly specific to sources likely to provide sports images. We seed the process with several popular query classes.

# 5. EVALUATION

To evaluate our approach for detecting source authority for web images, we amass a collection of over 10 million images. This data set is collected by issuing over 12,000 queries against a commercial web image search engine (in this case, Yahoo! image search) and collecting the top-1,000 returned images, which is the maximum number of results returned by such engines. In addition to the images, themselves, we also retain information about the sources of the images, such as the URL of the image and the URL of the web page that refers to the image itself. We further obtain the HTTP headers for each image in order to identify the 'Last-Modified' date for the image files.

## 5.1 Primary Sources

We hypothesize that the high-authority sources that we are identifying are actually often the originators of images, or the "primary sources." To investigate this, we offer some qualitative analysis of these sources. Figure 2 shows the top-ranked sources for a few classes of queries as predicted by our system. Within these results, we see a tendency for intuitively-identified primary sources to come up towards the top. In the "Athletes" category, we see "assets.espn.go.com" and "sportsillustrated.cnn.com," respectively the locations of images for the official ESPN and Sports Illustrated websites. For the "Actors" category, we see "us.movies1.yimg.com" and "mtv.com," which are the sources for images for Yahoo! Movies and MTV, respectively. Interestingly, for the "Pets" category, users on Flickr are providing the most authoritative content.

## 5.2 Relevance

Previous work [4] has established that the detection of near-duplicate images in image search results is a key cue for determining the relevance of images. In particular, images can be clustered into groups in which are all the images are near-duplicates of each other. The relative size of each cluster will reflect the number of times that the image has been copied, so more-frequently copied images can be ranked above less-frequently copied ones. A problem with this approach, like many other reranking techniques [1, 3, 5] is that it is required to be run over the results for every query. However, the process takes far too long to be suitable for use in an actual retrieval system.

We take 80% of our queries and construct citation networks and to yield rankings of websites. We then use these rankings against the search results for the remaining 20% of queries. We also apply an expensive duplicate-based reranking approach to the results of these test queries. We compare the relative rankings that would result from the duplicate reranking (the size of the duplicate clusters that each image would appear in) against the authority scores for the sources of each image. We find that these two values are highly correlated ($p << .001$), so we are able to approximate the results of a computationally expensive operation at very little run-time cost.

## 5.3 Fresh Images

There is a time lag between the creation of an image and the point at which it has become largely copied and redistributed across the web, so for a relatively new image, there would only be one or two instances of the image: not enough to warrant a high rank in reranking based approaches. Can

| Athletes | Movies | Games | Actors | Automotive | Pets |
|---|---|---|---|---|---|
| assets.espn.go.com | images.rottentomatoes.com | img.jeuxvideo.fr | us.movies1.yimg.com | img2.netcarshow.com | farm1.static.flickr.com |
| sportsillustrated.cnn.com | soundtrackcollector.com | tothegame.com | mtv.com | farm1.static.flickr.com | z.about.com |
| i.a.cnn.net | cinema.com | xboxmedia.ign.com | l.yimg.com | zercustoms.com | farm4.static.flickr.com |
| images.sportsline.com | moviegoods.com | gamershell.com | cinema.com | s2.desktopmachine.com | farm3.static.flickr.com |
| onlinesports.com | us.movies1.yimg.com | dignews.com | img2.timeinc.net | seriouswheels.com | lionking.org |
| canmag.com | z.about.com | gameslave.co.uk | images.zap2it.com | canadiandriver.com | eugenewei.com |
| firstshowing.net | kino.ural.ru | ps3media.ign.com | filmweb.no | mondeo.fordclubs.org | farm2.static.flickr.com |
| sportsmed.starwave.com | impawards.com | totalmortalkombat.com | img.timeinc.net | allcarwallpapers.com | icicom.up.pt |
| espn.go.com | dvdactive.com | pcmedia.ign.com | a69.g.akamai.net | farm4.static.flickr.com | arcatapet.com |
| espn-att.starwave.com | collider.com | xbox360media.ign.com | artistdirect.com | autogaleria.pl | snakesonablog.com |

**Figure 2: Top-ranked web sources for several classes of queries.**

we trust certain sources enough to believe that new images from the sources will still be likely to relevant? To investigate this question, we re-use the 80/20 training/testing split described in the previous evaluation. Here, we rank all of the images in the test set according to their "Last-Modified" date. We then take the 1% most-recent images and subsample two sets of 100 images: the first set are images from sites ranked to be the single most authoritative for their class and the second set is randomly sampled, regardless of site. We find that these fresh images are, on average, 71% relevant to the query, while the images from authoritative sites are relevant 95% of the time.

## 5.4 Citation in Social Media

We also conduct one further experiment (orthogonal to the above-described ones) to evaluate an application where the quotation action is more explicit. On Flickr, users can mark other users' photographs as "favorites," which is also effectively a citation. To experiment with this signal, we take a snapshot of Flickr as of April 2008: 25.5 million users, 2.5 billion photos, and 112 million instances of users marking photos as favorites. We construct a graph between users by building an edge from user A to user B if user A has marked one of user B's photos as a favorite. The edges are weighted by the total number of user B's photos that user A has favorited. 2.2 million users have marked a favorite or had one of their photographs marked as a favorite.

With a random walk over this graph of users, we can arrive at a listing of the most authoritative users: those most likely to be cited or favorited. We test this by ranking users according to their authority scores and gauging their propensity to produce highly interesting photographs (as determined by Flickr's "interestingness" algorithm). We gather the top 500 most-interesting photographs for the month following the end of our data set and find that more than 80% of the most-interesting photographs for that time period are produced by the top 2% most-quoted users. The top 10% produce nearly 100% of those photographs. Therefore, we can predict photographs of future interest based on the past frequency with which a particular photographer was cited.

# 6. CONCLUSIONS AND FUTURE WORK

We have proposed a new citation cue and a framework for judging the relevance of images based on the authority of the sources that provide the images. We have found our proposed approach is able to approximate relevance improvements that can be derived from other content-based post-query processing techniques without requiring significant computational resources at query time. The approach is also able to find brand new content that is relevant. We further apply the approach to photographs on a social photo sharing site and find that we are able to find users that often produce interesting content.

Thus far we have simply demonstrated the potential impact of source authority as derived through the proposed methods for ranking. Future work might deploy the source authority cue as a feature in a machine-learned-ranking application to fully utilize it in combination with any number of other factors.

# 7. REFERENCES


[1] W. H. Hsu, L. Kennedy, and S.-F. Chang. Video Search Reranking through Random Walk over Document-Level Context Graph. In *ACM Multimedia*, Augsburg, Germany, September 2007.

[2] Q. Jiang and Y. Zhang. SiteRank-Based Crawling Ordering Strategy for Search Engines. In *Proceedings of the 7th IEEE International Conference on Computer and Information Technology*, pages 259–263. IEEE Computer Society, 2007.

[3] Y. Jing and S. Baluja. PageRank for product image search. In *Conference on the World Wide Web*. ACM New York, NY, USA, 2008.

[4] L. Kennedy. *Advanced techniques for multimedia search: Leveraging cues from content and structure.* PhD thesis, Columbia University Graduate School of Arts and Sciences, 2009.

[5] J. Liu, W. Lai, X.-S. Hua, Y. Huang, and S. Li. Video search re-ranking via multi-graph propagation. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 208–217, New York, NY, USA, 2007. ACM.

[6] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Computer Science Department, Stanford University, 1998.

[8] E. Schmidt and J. Singh. SiteRank: Link-Based Relevance Computation for Persistent Search. Technical report, 745-06, Department of Computer Science, Princeton University, 2006.

[9] SIFT demo program, David Lowe. http://www.cs.ubc.ca/~lowe/keypoints/.

[10] J. Wu and K. Aberer. Using SiteRank in p2p information retrieval. Technical report, IC/2004/31, Swiss Federal Institute of Technology, Lausanne, Switzerland, 2004.