Integrated Intelligent Research (IIR)
International Journal of Computing Algorithm
Volume: 04 Issue: 01 June 2015 Pages:1-5
ISSN: 2278-2397

# An Emphasized Dual Similarity Measure Integration for Online Image Retrieval System Using SimRank

R.RajKumar[1], M.Krishnamurthy[2]
[1]PG Scholar, Department of CSE, KCG College of Technology, Chennai, Tamil Nadu, India
[2]Professor & Head ME-CSEKCG College of Technology, Chennai, Tamil Nadu, India
Email: rajcse4891@gmail.com, mkrish@kcgcollege.com

Abstract-In the real world scenario the use of image grows rapidly, the image rich network is the one that comprises of billions of images. The social media websites, such as Picasa, Flickr and Facebook comprises billions of end user posted images along with their annotations. Similarly the electronic commerce website such as Flipkart, Myntra and Amazon are also furnished with product related images. In this paper, we introduce how to perform efficient and optimum information retrieval in online image rich system. We propose a Mok-SimRank to compute link-based similarity and a dual similarity integration algorithm for both link and content based similarity. Experimental results on online electronic commerce site show that our approach is significantly better than traditional methods in terms of relevance.

Keywords – Dual similarity, Image retrieval, simrank

## I. INTRODUCTION

Social multimedia websites such as Flickr, Facebook, Picasa, etc., are popular around the world, with over billions of photos, images posted, shared by users. The Internet electronic commerce website such as Flipkart is also furnished with tremendous amounts of product-related images. This kind of network can be treated as image loaded network [10].Fig.1 represents the Flipkart information network of product images, categories, and consumer tags. Information retrieval [9][5] in such large image loaded information networks is a very useful but also very complex task, because there exists a lot of information such as text, image feature, user, group, etc. Document retrieval systems find information to given criteria by matching text records (documents) against user keyword, as opposed to expert systems that answer questions by inferring over a logical knowledge database. A document retrieval system consists of a database of documents, a classification algorithm to build a full text index, and a user interface to access the database. A document retrieval system has two main tasks:

1) Find relevant documents to user queries
2) Evaluate the matching results and sort them according to relevance, using algorithms such as PageRank.

Content-based retrieval means that the search analyses the contents of the image rather than the metadata such as keywords, tags, or descriptions associated with the image.The term "content" in this context might refer to colours, shapes, textures, or any other information that can be derived from the image itself. Most commercial image search engines use textual similarity to return semantically relevant images and then use visual similarity to search for visually relevant images. However, existing works cannot handle the link structure. We are using the Mok-SimRank algorithm to estimate the link structure similarity. When consider the images in the network, image similarity can actually also be estimated by content features, such as RGB histogram and SIFT. Then, we propose algorithm DSIL to provide a novel way of integrating both link and content information.

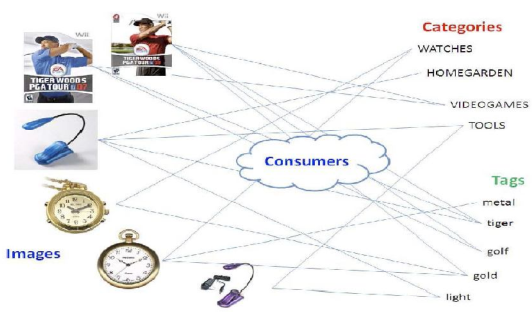


Figure 1.Information network for Flip kart, connected by products, usertags, and categories.

## II. RELATED WORK

Digital imagery is becoming important component of computer and telecommunication usage. However the increasing use of imagery is causing severe problems, because the technology for organizing and searching images based on their content is still in its infancy. Currently the standard approach to searching image and video is to create text annotations that describe the content of the image and then enter these textual annotations into a standard database. The image themselves are not really part of the database; they are only referenced by text strings or pointers. These annotations must be entered manually with great tedium and prohibitive cost.

Differently from pure CBIR [7] systems Visual Rank [8] retains the commonly used text query interface and utilizes the visual similarities within the entire set of images for image selection. This approach complements pure CBIR systems in several ways: 1) Text is still the most familiar and, often, the only query medium for commercial search engine users, 2) VisualRank can be effectively used in combination with other CBIR systems by generating a more relevant and diverse set of initial results, which often results in a better starting point for pure CBIR systems, and 3) there are real-world usage scenarios beyond “traditional” image search where image queries are not feasible. In many uses, we need to select a very small set of images to show from potentially millions of images. Unlike

International Journal of Computing Algorithm
Volume: 04 Issue: 01 June 2015 Pages:1-5
ISSN: 2278-2397

ranking, the goal is not to reorder the full set of images but to select only the “best” ones to show.

In “Image Retrieval: Current Techniques, Promising Directions, and Open Issues, by Y. Rui, T.S.Haung”State that the image retrieval from different views, one is purely text based and another one is based on the visual features. Most popular framework of image retrieval then was to first annotate the images by text and then use text-based database management systems (DBMS) to perform image retrieval.Image Meta search is a type of search engine specialised on finding pictures, images, animations etc. Like the text search, image search is an information retrieval system designed to help to find information on the Internet and it allows the user to look for images etc. using keywords or search phrases and to receive a set of thumbnail images, sorted by relevancy. A common misunderstanding when it comes to image search is that the technology is based on detecting information in the image itself. But most image search works as other search engines. The metadata of the image is indexed and stored in a large database and when a search query is performed the image search engine looks up the index, and queries are matched with the stored information. The results are presented in order of relevancy.

The usefulness of an image search engine depends on the relevance of the results it returns, and the ranking algorithms are one of the keys to becoming a big player. Some search engines can automatically identify a limited range of visual content, e.g. faces, trees, sky, buildings, flowers, colours etc. This can be used alone, as in content-based image retrieval, or to augment metadata in an image search. When performing a search the user receives a set of thumbnail images, sorted by relevancy. Each thumbnail is a link back to the original web site where that image is located. Using an advanced search option the user can typically adjust the search criteria to fit their own needs, choosing to search only images or animations, colour or black and white, and setting preferences on image size.

## III. CONTENT BASED IMAGE SIMILARITIES

In this section, we describe the common feature extraction process and the frequently used feature representation form, the so-called *featuresignatures*. The extraction of data object features and their aggregation aim at digitizing and compactly storing the data objects inherent properties. In the feature extraction step, each data object is mapped into a set of features in an appropriate feature space *FS*. In the field of content-based image retrieval, the feature space frequently comprises position, colour, or texture dimensions where each image pixel is mapped to a single feature in the corresponding feature space. In this way, the content of each data object is exhibited via its feature distribution in the feature space.

Using visual similarity only may lead to irrelevant results. Fig. 2 shows one pair of Flickr images that have similar content similarity estimated from the low-level feature, but with different semantic meaning. Direct use of link information solely based on human annotations[4] may also lead to unsatisfying results if the annotation is wrong, too general, or incomplete. In addition, if the image does not link to any object in the information network, then only based on link information cannot work. Fig.3 shows several examples that are all linked to tag “flower” but they are not visually similar.

### *A. Similarity Metric*

Image similarity can be estimated from image content features , such as colour histogram, edge histogram , Colour Correlogram , CEDD , GIST, texture features , Gabor features , shape[3] and SIFT .Normalize feature F RD, where D is the number ofdimensions in the feature space, to be of unit length: for anyfd, the value of feature F on dimension d (d = 1; . . .; D),divide it by the sum of values on all dimensions

$$f^d = f^d_{orig} \Big/ \sum_{d=1}^{D} f^d_{orig}.$$

The chi-square test statistic distance between two feature vectors Fi and Fj is defined as:

$$\chi_{ij} = \chi(F_i, F_j) = \sum_{d=1}^{D} c^d_{ij}$$

$$= \frac{1}{2} \sum_{d=1}^{D} \frac{(f_i^d - f_j^d)^2}{f_i^d + f_j^d}.$$

Figure:2 Images with high visual similarity, but low semantic similarity.

Figure :3 Images annotated by the tag “flower,” but with low visualsimilarity.

## IV. LINK- BASED SIMILARITY

SimRank is a general similarity measure, based on a simple and intuitive graph-theoretic model. SimRank is applicable in any domain with object-to- object relationships, that measures similarity of the structural context in which objects occur, based on their relationships with other objects. Effectively, SimRank is a measure that says "two objects are considered to be similar if they are referenced by similar objects."SimRank is a link-based similarity measure, and builds on the approach of previously existing link-based measures. SimRank is based on both a clear human intuition and a solid theoretical background. Similarly to PageRank [6], SimRank is defined recursively with respect to “random surfer" model and iscomputed iteratively. Unlike the similarity measures that require human-built hierarchies, SimRank is applicable to any domain with object-to-object relationships, including the Web

Nevertheless, existing work on SimRank lacks two important issues. Firstly, although SimRank iterative similarity scores are known to converge, a real-life computationnaturally involves performing a finite number of iterations. Secondly, optimization issue of SimRank computation is not the primary focus of the original SimRank proposal.For a node $v$ in a directed graph, we denote by $I(v)$ and $O(v)$ the set of in-neighbours and out-neighboursof $v$, respectively. Individual in-neighboursare denoted as $I_i(v)$, for $1 \leq i \leq |I(v)|$, and individual out-neighbours are denoted as $O_i(v)$, for $1 \leq i \leq |O(v)|$.

Let us denote the similarity between objects $a$ and $b$ by $s(a,b) \in [0,1]$. Following the earlier motivation, a recursive equation is written for $s(a,b)$. If $a = b$ then $s(a,b)$ is defined to be $1$. Otherwise,

$$s(a,b) = \frac{C}{|I(a)|\,|I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$$

Where $C$ is a constant between $0$ and $1$. A slight technicality here is that either $a$ or $b$ may not have any in-neighbours. Since there is no way to infer any similarity between $a$ and $b$ in this case, similarity is set to $s(a,b) = 0$, so the summation in the above equation is defined to be $0$ when $I(a) = \emptyset$ or $I(b) = \emptyset$.

HMok-SimRank:Mok-SimRank can be extended to work for a weighted heterogeneous information network. To demonstrate this method, we take the image-rich information network from Flickr as an example. Similar images are link to similar groups and tags, so we define the link-based semantic similarity between images. Weight can be set manually or automatically. Take FlipKart as anexample, the tag frequency represents the number of users who think the tag is relevant to the product. So we can use the tag frequency (or log value) as weight for the link between product image and tag. The group similarity is computed via the similarity of the images and tags they link to. The tag similarity is calculated via the similarity of the images and groups they link to.

The ultimate work of this paper is as follows:

1.We propose HMok-SimRank to efficiently compute weighted link-based similarity in weighted heterogeneous image-rich information networks. The method is faster than heterogeneous SimRank and K-SimRank.

2. We propose a dual approach based on link based and content based for similarity measures.

3. We propose the algorithm DSIL to provide a novel way of reinforcement style integrating with feature weighting learning for similarity/relevance computation in weighted heterogeneous image-rich information network.

## V. SYSTEM DESIGN

The overview of the proposed architecture for the dual similarity measure starts with the input of the search query which compares with the database for the link and content based similarity. From here the match type is seen and the requested web search done which gives the relevant search result as shown in Fig 4..In this case the system architecture shows how the flow is processing in this work. In image content-based retrieval, most methods and systems compute image similarity based on image content features. Hybrid approach combine text features and image content features together. Most commercial image search engines use textual similarity to return semantically relevant images and then use visual similarity to search for visually relevant images. Integration-based approaches use linear or nonlinear combination of the textual and visual features.

### A. *Link -Based Similarity*

SimRank [2] is one of the most popular link-based algorithms for evaluating similarity between nodes in information networks. It computes node similarity based on the idea that “two nodes are similar if they are linked by similar nodes in the network.” Inspirit of PageRank, SimRank computes the similarity between each pair of nodes in an iterative fashion with a theoretical guarantee of the convergence.Similar images are likely to link to similar tags and groups, so we define the link-based semantic similarity between images as combination of similarity of group and similarity of tags. It is defined as follows This module iteratively calculate the similarity between image pairs, similarity between group pairs of images and similarity between tag pairs of image until the convergence is reached.

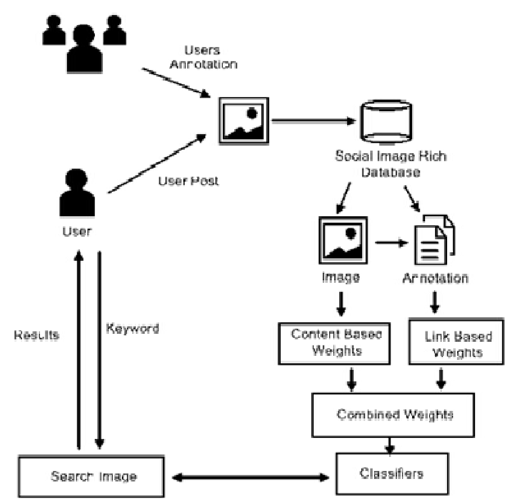


Figure:4 System Design

### B. *Content-Based Similarity*

Features such as colour histogram, edge histogram, ColourCorrelogram, CEDD, GIST, Texture features, Gabor features shape and SIFT are used for content based image retrieval. It represent an image as a point in a D-dimension feature space with either a single type of feature or a combination of multiple Image similarity can be estimated from image content types of features. Tang at el. proposed strategies to integrate both local and global features. If the integrated feature space has fixed number of dimensions, our approach is also applicable. The image vector information is

International Journal of Computing Algorithm
Volume: 04 Issue: 01 June 2015 Pages:1-5
ISSN: 2278-2397

extracted from the image content based on colour and histogram and this vector information is used by the cosine similarity function to measure the similarity. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude.

**Algorithm: Content-Based Similarity Measures**

Input: Social information So
Output: Requested images.
Method:

Step 1. Extract the social information So (image I)

Step 2. Extract image vector Ķ from images. This can be done by using RGB colour histogram.

Step3. Calculate histogram by using following function: hist(R, G, B) = ϴ

Step4. Calculate Cos (ϴ) = 0, 1,-1 similarity function

**Algorithm: Link-Based Similarity Measures**

Input: Social information So
Output: Requested images.
Method:

Step 1.Extract the social information So (annotation ā)

Step 2. Perform normalization from ā for keyword extraction.

Step 3.Stop word removal technique can be used to extract the keyword.

Step 4. Find keyword frequency fi

Using content similarity only may lead to unsatisfying results. It shows one pair of images that have similar content similarity estimated from the low-level feature, but with different semantic meaning. Direct use of link information solely based on human annotations may also lead to unsatisfying results if the annotation is wrong, too general, or incomplete. In addition, if the image does not link to any object in the information network, then only based on link information cannot work Shows several examples that are all linked to tag "flower" but they are not visually similar. This paper presents a novel algorithm to integrate link-based and content-based similarities: First perform HMok-SimRank to compute the link-based similarities and second perform feature learning considering the link-based similarity to update the feature weights, and then update the node similarities based on the new content similarity.

**Algorithm: Dual Similarity Integration (DSI)**

Input: G, the image-rich information network.

1. Construct kd-tree (or LSH and cv-tree index) over the image features;

2. Find top k similar candidates of each object;

3. Initialize similarity scores;

4. Iterate {

5. Calculate the link similarity for image pairs via HMok-SimRank;

6. Perform feature learning to update W using either global local feature learning;

7. Update the new image similarities

8. Compute link-based similarity for all group and tag pairs viaHMok-SimRank;

9.} until converge or stop criteria satisfied.

Output: S, Similarity scores.

## VI. EXPERIMENTS

### A. Data Sets

The experiments were conducted on the synthetic data sets generated based on an electronic commerce website. The data set is created by downloading product images and related metadata information, such as category, tags, and title. Product category is treated as group. For image feature extraction, we extracted CEDD, which is a compact descriptor that considers both colour and edge features. In theoretical guidance, it has shown good performance compared with many traditional features.

### B. Result Analysis

The Figure 5 is snapshotted from the famous e-commerce website FlipKart. It demonstrates that for the given keyword "iPhone 6", it retrieves the images of the given keyword with some irrelevancy. Therefore Images annotated by the tag "iPhone 6," but with low visual similarity. The figure 6 demonstrates the proposed DSIL algorithm, the top K-most image is retrieved for the specified keyword; the system obtains the best results in terms of the relevance for both semantic and visual appearances. The object is tagged with "iPhone, invisible shield, accessories" and belongs to category "Smart phone." Again, DSIL obtains the best results in terms of the relevance for both semantic and visual appearances. Our experiments also show good performance of our algorithm to find similar groups and relevant tags. We can use such tag similarity to help find more relevant images for a keyword query.

Figure:5 .The query image from FlipKart website


Figure: 6 The query image from our Dual Similarity Integration System

The images retrieved by the electronic commerce website “FlipKart” for the given keyword iphone6, clearly demonstrates that the existing system has some irrelevancy associated with-it.The top K-most image is retrieved for the specified keyword; the system obtains the best results in terms of the relevance for both semantic and visual appearances. The object is tagged with “iPhone, invisible shield, accessories” and belongs to category “Smart phone.”

C. *Performance Evaluation Method.*

The mean average precision is used to measure the retrieval performance of the various algorithms. For each image in the synthetic data set, we gather a ranking list of relevant images computed by each algorithm and compute the average precision based on the approximate ground truth before removing tags. The final MAP score for each algorithm is estimated as the mean average precision ofeach image. There is no need of training data set, i.e. all the algorithms are unsupervised one. Figure 7 show the result on Amazon data, respectively. We can observe that link-based similarity performs better than text-based similarity; VLWC achieves better performance than traditional algorithms by linearly combining visual and link information together. Algorithm DSIL further improves the performance by introducing a novel way of integrating content and link information.

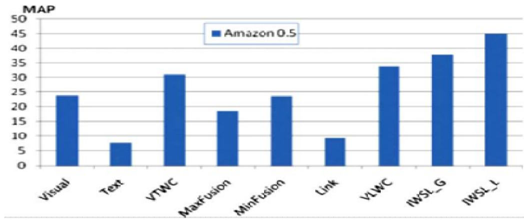


Figure:7. MAP of the algorithms on Amazon data. X-axis denotes the algorithms. Y -axis denotes the MAP (percent).

## VII. CONCLUSIONS

This paper presents a novel and efficient way of finding similar objects .Our major contributions are as follows:We propose HMOK-SimRank to efficiently compute weighted link-based similarity. This method is much faster than heterogeneous SimRank and K-SimRank.We propose the algorithm DSIL to provide a novel way of integrating link and content similaritiesWe conducted experiments on electronic commerce networks. The results have shown that our algorithm achieves better retrieval performance.Thus a new product search system to find both visually similar and semantically relevant products based on our algorithms HMOK-SimRank and Feature Learning, is proven to be efficient than theexistingsystem.As a *future work* in retrieving large scale web image with multiple visual concepts, it is challenging for Parts-based probabilistic model to outperform. So the work should extend its feature over pairwise visual similarity among images.

## REFERENCES

[1] Botterill, Mills, Green,”Speeded-Up Bag-Of-Words Algorithm for Robot Localization through Scene Recognition”, IEEE Conference on Image and Vision Computing,pp.1-6,Year 2008.
[2] Dmitry Lizorkin, PavelVelikhov, Maxim Grinev, Denis Turdakov, “Accuracy Estimate and Optimization Techniques for SimRank Computation”, The VLDB Journal, Volume 19, Issue 1, pp. 45-66, Year 2010.
[3] Greg Mori, Serge Belongie, Jitendra Malik, “Efficient Shape Matching Using Shape Context”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 27, Issue 11, p7p.1832-1837,Year 2005.
[4] Jinhui Tang, Haojie Li, Guo Jun Qi, Tat Seng Chua, “Image Annotation by Graph-Based Inference with Integrated Multiple/Single Instance Representations”, IEEE Transactions on Multimedia, Volume: 12, Issue: 2, pp131-141, Year 2010.
[5] KalervoJarvelin and JaanaKekalainen, “IR Evaluation Methods for Retrieving Highly Relevant Documents”, Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41-48,Year 2008.
[6] Larry Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, “The Pagerank Citation Ranking: Bringing Order to the Web”, technical report, Stanford University,Year 2008.
[7] Raghu Krishnapuram, SwarupMedasani, Sung Hwan Jung , Young Sik Choi, Rajesh Balasubramaniam, “Content-based image retrieval based on a fuzzy approach”,IEEE Transactions on Knowledge and Data Engineering Volume:16 Issue:10, pp. 1185-1199,Year 2010.
[8] ShumeetBaluja, Yushi Jing, “VisualRank: Applying PageRank to Large-Scale Image Search”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 30, Issue: 11, pp.1877-1890, Year 2008.
[9] Siddiquie, Feris, Davis, “Image Ranking and Retrieval based on Multi-Attribute Queries”, IEEE Conference on Computer Vision and Pattern Recognition, pp.801-808, Year 2011.
[10] Xin Jin, JieboLuo, Jie Yu, Gang Wang, Dhiraj Joshi and JiaweiHan,”Reinforced Similarity Integration in Image-Rich Information Networks”, IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 2, pp.448-460,Year 2013.