IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 2, Apr-May, 2014
ISSN: 2320 - 8791
www.ijreat.org

A Hybrid Similarity Measures For Integration In Image-Rich Information Networks

V.Abinaya[1], Mr.A.R.Ashok Kumar[2], Prof.D.Durai Kumar[3]

[1]M.Tech, Department of Information Technology, Ganadipathy Tulsi's Jain Engineering college, Vellore, Tamil Nadu, India,

[2]Assistant Professor, Department of IT, Ganadipathy Tulsi's Jain Engineering College, Vellore, Tamil Nadu, India,

[3]Head Of The Department IT, Ganadipathy Tulsi's Jain Engineering College, Vellore, Tamil Nadu, India,

Abstract

An image-rich information network is a social media website with billions of images uploaded by users which are associated with information about owner, consumer, producer, annotations, and comments. In this paper, the problem of information retrieval in an image- rich information network and recommendation in such networks is addressed. We propose a combined approach which measures the similarity based on both link based and Content based. The link similarity depends upon the social network information like tags, groups and human annotation over the images. Content based similarity considers the image content properties color histogram etc., for measuring the similarity. The combined score of these measures is used to integrate the social resources and helps to classify the images in image-rich information networks.

Keywords- image-rich, combined score, similarity, classification

1.Introduction

Social multimedia sharing and websites for hosting, such as Flickr, Facebook, YouTube, Picasa, etc., are popular around the world, with over billions of photos, images uploaded, shared by users. Most Popular Internet commerce website such as Amazon are also furnished with tremendous amounts of product-related images. In this way, many images in such social networks are accompanied by information about owner, consumer, producer, annotations and comments. These kind of image related information can be modeled as heterogeneous image-rich information networks. Feature (content) extraction is the basis of content-based image retrieval. In this case, features may include both text-based features (key words, annotations)and visual features (color, texture, shape, faces). In text-based retrieval estimating the similarity of the words in the context is useful for returning more relevant images. Word Net manually groups words into synonym sets, Google Distance [1] computes word similarity by co-occurrence in search results. Flickr Distance [2] considers visual relationship in depth manner. In image content-based retrieval, most methods (such as Google's VisualRank [3]) and systems [4], [5], [6], [7], [8] compute image similarity based on the image content features. Hybrid similarity approach combine text features and image content features together [9]. 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. But still, existing works cannot handle the link structure in efficient manner. In this paper, an image-rich information network model where the similarities between same type of nodes and different types of nodes can be better estimated based on the mutual impact under the network structure. Among all algorithms that compute object similarity in information networks, SimRank is one of the most popular algorithm, but it is very expensive to calculate and the similarity is only based on the link information. When consider images in the network, image similarity can actually be judged by content features, such as RGB histogram and SIFT. In this paper, propose an efficient approach called Mok-SimRank to significantly improve the speed of k-Sim-Rank, and introduce its extension HMok-SimRank to work on weighted heterogeneous information networks. Image similarity can be estimated from image content features, such as color histogram, edge histogram , Color Correlogram , CEDD , GIST, texture features, Gabor features, shape and SIFT.

HMok-SimRank:Mok-SimRank can be extended to work for a weighted heterogeneous information network. To explain 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 Amazon as an example, 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.
Major work of this paper 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 combined approach based on link based and content based for similarity measures.
3. We propose the algorithm IWSL to provide a novel way of reinforcement style integrating with feature weighting learning for similarity/relevance computation in weighted heterogeneous image-rich information network.

2. Related Work

Direct use of link information solely based on human annotations may also lead to unsatisfying results if the annotation is meaningless, too general, or inadequate. In additionally, if the image does not link to any object in the information network, then link information cannot work. Using content similarity only may lead to unsatisfying results.

In [10] B.Babenko,et al proposed "Similarity Functions for Categorization:From Monolithic to Category Specific," gives detail information about Similarity metrics that are learned from labeled training data can be advantageous in terms of performance and efficiency. For the purpose of similarity categorization two approaches have been defined:
1.Learning a global or "monolithic similarity metric

2.Learning a similarity metric per category

There are several advantages to training a monolithic metric.Such a matric can be used in a nearest neighbor classifier, which can lent itself to efficient classification. Furthermore, the representation is the same for all the data.This is convenient because the matric can easily generalize to novel categories.

In [7] V.N. Gudivada, et al proposed Content-Based Image Retrieval Systems, content-based image retrieval paradigm. There are three fundamental bases for content-based image retrieval, 1. visual feature extraction, 2. multidimensional indexing, 3. retrieval system design. They also give details about silmilarity-based retrieval. One of the main difference between image retrieval system and a database system is the former's ability to rank-order database images by the degree of similarity with the query image(i.e.similarity based retrieval). Database system typically process queries based on exact match. A theoretical framework for similarity-based retrieval should be developed.

"IRIN: Image Retrieval in Image-Rich Information Networks," [11] gives detail information about efficient approach called Mok-SimRank to significantly improve the speed of SimRank, and propose an algorithm called k-SimRank to consider both link and content information by seamlessly integrating reinforcement learning with feature learning. In an image-rich information network, similar images are likely to link to similar groups and tags, so we define the linkbased semantic similarity between images. The group similarity is computed via the similarity of the images and tags they link to, and the tag similarity is calculated via the similarity of the images and groups they link to.
"Accuracy Estimate and Optimization Techniques for SimRank ," [12] defines the similarity measures for SimRank can be considered as one of the auspicious ones, because of the following reasons. 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. The following optimization techniques are defined:

- Selecting Essential Node Pairs
- Partial Sums
- Threshold-Seived Similarity

In [8] "Image Retrieval: Current Techniques, Promising Directions, and Open Issues, by Y. Rui, T.S.Haung "State that the image retrieval from different angles, one is text-based and another one is visual-based. 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.

3. Motivation

Conducting information retrieval in a large image-rich information networks is a very useful but also very difficult task, because it contains more number of information like text, image feature, user, group, and most importantly the network structure. The objective of this paper is to extract the similarity between image rich social resources using content and social annotation of online social media.

4. System Design

4.1 Architecture

The Mok-SimRank algorithm measures the link based similarity by finding the similarity of group and similarity of tags separately then adding them. Cosine similarity is used to calculate the content similarity between the two images Finally apply integration technique to combine the link based and content based similarity measure. Then classifies the images using classifier. These measures are used by our recommendation system to suggest social image resource to the users.

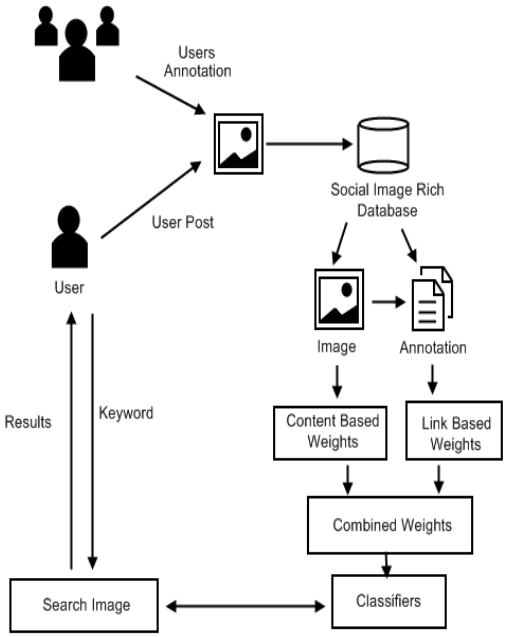

Figure 1. System Architecture

4.2 Modules

4.2.1 Social Community Development

This module is used to construct the basic Social multimedia platform similar to the flicker through which the image and image related data is collected and preprocesses step to construct the required data model for our Similarity Integration. This module contains user registration and login process, image posting, sharing and human based annotation like tagging of images and image preprocessing step.

4.2.2 Link-Based Similarity

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

$$S_{m+1}(e,e') = \alpha_I S_m^G(e,e') + \beta_I S_m^T(e,e'),$$

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.

4.2.3 Content-Based Image Similarity

The image vector information is extracted from the image content based on color 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.

4.2.4 Similarities Integration

We present 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.

4.2.5 Formation of Clustering

The final weight calculated after integration of both link based similarity and content similarity is used to classify the images based on their similarity. From this clustered information users can retrieve their needed images by using keyword.

4.3 Algorithms

4.3.1 Hmok-Simrank Algorithm

Hmok-Simrank algorithm is a combination of both link-based and content-based similarity measures. Following algorithms explain about these two similarity measures.

Algorithm 1-Link-based similarity measures

Input: social information **So**
1.Extract the social information **So** (annotation **ā**)
2.Perform normalization from **ā** for keyword extraction.
3.Stop word removal technique can be used to extract the keyword.
4.Find keyword frequency **fi**

5.Top k results from **fi** stored in database.
Output: requested images.
Algorithm 2-Content-based similarity measures
Input: social information **So**
1.Extract the social information **So** (image **I**)
2.Extract image vector **Ƙ** from images. This can be done by using RGB color histogram.
3.Calculate histogram by using following function:
$\sum$ hist(R,G,B) = **Ɵ**
4.Calculate Cos(**Ɵ**) **= 0,1,-1** similarity function
Output: requested images.

4.3.2 Feature Learning Algorithm

To build a bridge between the content and semantics, we learn a weighting vector for the feature space to force the weighted content-based similarity to be somehow consistent with the semantic link-based similarity.We consider two types of approaches: Global Feature Learning (GFL), Local Feature Learning (LFL),

Global Feature Learning :

If the tags (or groups) of an image are incomplete (0 or very few) and thus cannot fully describe its semantic meaning, the link-based similarity becomes unreliable. In order to overcome this factor, we introduce the confidence (or importance) and define it as a function of the number of linked annotations (including both tags and groups) to the image. After global feature learning, update the image similarity as a combination of content-based and link-based similarity.

Local Feature Learning

The problem of global feature learning is that using a global feature weighting for all images may be too general. Different images may belong to different semantic topics and thus need different weightings to capture their specific important features. Therefore, we can perform LFL to find a specific feature weight for image.

4.3.3 Integration Algorithm
Input: G, the image-rich information network.
1.Construct kd-tree over the image features;
2.Find top k (or e-range) 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 = W*m+1 using either global or local feature learning;
7.(Optional) Search for new top k similar image candidates based on the new similarity weighting;
8.Update the new image similarities Sm+1 (i,i') global or local feature learning;
9.Compute link-based similarity for all group and tag pairs via Hmok-SimRank;
10.} until converge or stop criteria satisfied.
Output: S, pair-wise node similarity scores.

5. Conclusions

This paper presents a novel and efficient way of finding similar objects (such as photos and products) by modeling major social sharing and e-commerce websites as image-rich information networks. We propose HMok-SimRank to efficiently compute weighted link-based similarity in weighted heterogeneous image-rich information networks. The method is much faster than heterogeneous SimRank and K-SimRank. We have implemented a new search and recommendation system to find both visually similar and semantically relevant products based on our algorithm.

6. References

[1] R.L. Cilibrasi and P.M.B. Vitanyi, "The Google Similarity Distance," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 3, pp. 370-383, Mar. 2007.
[2] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li, "Flickr Distance," Proc. 16th ACM Int'L conf.
[3] Y. Jing and S. Baluja, "VisualRank: Applying Pagerank to LargeScale Image Search," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 30, no. 11, pp. 1877-1890, Nov. 2008.
[4] R.C. Veltkamp and M. Tanase, "Content-Based Image Retrieval Systems: A Survey," technical report, Dept. of Computing Science, Utrecht Univ., 2002.
[5] H. Tamura and N. Yokoya, "Image Database Systems: A Survey," Pattern Recognition, vol. 17, no. 1, pp. 29-43, 1984.
[6] W.I. Grosky, "Multimedia Information Systems," IEEE MultiMedia, vol. 1, no. 1, pp. 12-24, Spring, 1994.
[7] V.N. Gudivada and V.V. Raghavan, "Content-Based Image Retrieval Systems," Computer, vol. 28, no. 9, pp. 18-22, Sept. 1995.
[8] Y. Rui, T.S. Huang, and S.-F. Chang, "Image Retrieval: Current Techniques, Promising Directions, and Open Issues," J. Visual Comm. and Image Representation, vol. 10, no. 1, pp. 39-62, 1999.
[9] R. Datta, D. Joshi, J. Li, and J.Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," ACM Computing Surveys, vol. 40, no. 2, pp. 1-60, Apr. 2008.
[10] B. Babenko, S. Branson, and S. Belongie, "Similarity Functions for Categorization: From Monolithic to Category Specific," Proc. IEEE 12th Int'l Conf. Computer Vision (ICCV '09), 2009.
[11] X. Jin, J. Luo, J. Yu, G. Wang, D. Joshi, and J. Han, "iRIN: Image Retrieval in Image-Rich Information Networks," Proc. 19th Int'l Conf. World Wide Web (WWW '10), pp. 1261-1264, 2010.
[12] D. Lizorkin, P. Velikhov, M. Grinev, and D. Turdakov, "Accuracy Estimate and Optimization Techniques for Simrank Computation," VLDB Endowment, vol. 1, no. 1, pp. 422-433, 2008.