Clustering Web Images using Association Rules, Interestingness Measures, and Hypergraph Partitions

Hassan H. Malik
Department of Computer Science
Columbia University
New York, NY 10027
+1-650-283-5054
hhm2104@columbia.edu

John R. Kender
Department of Computer Science
Columbia University
New York, NY 10027
+1-212-939-7115
kender@cs.columbia.edu

ABSTRACT

This paper presents a new approach to cluster web images. Images are first processed to extract signal features such as color in HSV format and quantized orientation. Web pages referring to these images are processed to extract textual features (keywords) and feature reduction techniques such as stemming, stop word elimination, and Zipf's law are applied. All visual and textual features are used to generate association rules. Hypergraphs are generated from these rules, with features used as vertices and discovered associations as hyperedges. Twenty-two objective "interestingness" measures are evaluated on their ability to prune non-interesting rules and to assign weights to hyperedges. Then a hypergraph partitioning algorithm is used to generate clusters of features, and a simple scoring function is used to assign images to clusters. A tree-distance-based evaluation measure is used to evaluate the quality of image clustering with respect to manually generated ground truth.

Our experiments indicate that combining textual and content-based features results in better clustering as compared to signal-only or text-only approaches. Online steps are done in real-time, which makes this approach practical for web images. Furthermore, we demonstrate that statistical interestingness measures such as Correlation Coefficient, Laplace, Kappa and J-Measure result in better clustering compared to traditional association rule interestingness measures such as Support and Confidence.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *Data mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Clustering*.

General Terms

Performance, Experimentation.

Keywords: Web image clustering, feature selection, web mining, association rules, interestingness measures, hypergraphs.

1. INTRODUCTION

1.1 Visual and Textual Mining

The last two decades have seen significant research in the field of data mining, resulting in a number of successful techniques such as finding associations between data items by mining association rules [1]. These techniques are proven to be very useful in symbolic and structured domains such as market basket analysis. Limited research, however, has been conducted to apply these techniques on non-structured, signal-based domains like images [12, 25, 27].

Unlike structured data where features such as keywords and alphanumeric values can be easily identified and extracted, images contain implicit features and patterns that are not straightforward to identify and extract [17]. The fundamental challenge in image mining is to determine how low-level pixel representations can be efficiently and effectively processed to identify these high-level patterns [17]. Once identified, these patterns could be used in a variety of applications.

Clustering is one such application that uses features to organize data in a number of groups called clusters. Two major approaches exist to cluster images: content-based and text-based. Content-based clustering is normally used by the image analysis and computer vision communities and focuses on exploiting low-level signal features like color, shape, and texture to cluster images, while text-based clustering is normally used by the web mining and information retrieval communities. A common perception exists in web and information retrieval communities [15] that content-based features are computationally expensive to extract and hence infeasible for the web domain. However, some features such as color and orientation can be extracted in linear time. Furthermore, applying simple techniques like image scaling can further reduce computational requirements.

In contrast, the availability of reasonable textual information is not always guaranteed. A large number of images on the web either do not have any textual information associated to them, or the associated textual information does not provide much information about the image (i.e. insufficient to disambiguate from other images that belong to different semantic categories but share some keywords). Text-only clustering techniques are very likely to assign such images to wrong clusters, resulting in low-quality clustering. Similarly, unless very sophisticated and computationally intensive techniques are used to capture semantics, signal-only clustering techniques are also likely to produce low quality clusters. We show that using a combination of textual and simple signal features results in better clustering as compared to clustering solely based on text or signal-only features.

ICWE'06, July 11–14, 2006, Palo Alto, California, USA.
ACM 1-59593-352-2/06/0007.

1.2 A Novel Approach

Hypergraphs are proven useful in data mining and high-dimensional document clustering problems [13, 14]. In a typical hypergraph, each vertex represents a dimension and each hyperedge represents affinity (or relationship) between two or more vertices. Weights assigned to vertices indicate importance of these vertices and weights assigned to hyperedges indicate the strength of the relationship between vertices connected by a hyperedge. In this paper, we first extract signal and text features from images, calculate their frequencies, and apply well-known dimensionality reduction techniques such as stemming, stop word elimination, and Zipf's law to prune non-interesting features. The remaining features are used to generate association rules.

Similar to [14], we use features as hypergraph vertices and all association rules between a set of vertices to generate hyperedges. In the last decade or so, various researches [6, 22] questioned the usefulness of Support and Confidence as association rule interestingness measures and have proposed various alternates. Unfortunately, researchers comparing interestingness measures [16, 32] do not agree on any single domain-independent objective measure. Considering this, we compared twenty-two objective interestingness measures to assign weights to hyperedges, rather then using Confidence [14].

Once the association rule hypergraph is available, we apply a widely used hypergraph partitioning algorithm hMETIS [18] to obtain partitions (or clusters) of features. Images are assigned to these clusters using a simple scoring function. This clustering method eliminates the need of calculating image distances or similarities against other images. Finally, we use a tree-distance-based evaluation measure to evaluate the quality of the resulting image clusters with respect to manually generated ground truth.

Most of the steps in this approach, including feature extraction, reduction, rule generation, feature hypergraph generation, and hypergraph partitioning can be performed offline. Assignment of images to clusters is the only real-time step, which is computationally inexpensive.

2. RELATED WORK

2.1 Web Image Clustering

There have been several web image clustering and categorization approaches proposed in recent years. We discuss only a few representative approaches here. Lienhart and Hartmann [21] use signal-only features to categorize web images. Images are divided into photo-like images, and graphical images. Photo-like images are further divided into photos and artificial photo-like images; graphical images are further divided into slides, cartoons, and other images. This approach produces coarse categories containing too many images. Although syntactically meaningful, the resulting clusters are likely to contain images that are not semantically related. In contrast, ImageSeer [15] uses the VIPS algorithm [7] to segment web pages into several semantic blocks. These blocks are further used to extract surrounding text of web images. Page-to-block, block-to-image, and block-to-page relationships are obtained using the link structure and page layout analysis, and an image graph is constructed. Techniques from spectral graph theory and Markov chain theory are applied for image ranking, clustering, and embedding. Like any other text-only approach, this approach is likely to assign images with insufficient textual information to wrong clusters.

2.2 Association Rules and Association Rule "Interestingness" Measures

The problem of mining association rules was first introduced in [1]. If $I = \{i_1, i_2, \ldots, i_m\}$ is a set of literals, called items and D is a set of transactions, an *association rule* is an implication of the form $X \rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = 0$. The rule $X \rightarrow Y$ holds in the transaction set D with *Confidence* c if 100c% of transactions in D that contain X also contains Y. The rule $X \rightarrow Y$ has *Support* s in the transaction set D if 100s% of transactions in D contain $X \cup Y$. Although a number of algorithms are proposed improving various aspects of association rule mining [5, 9, 10], Apriori [2] remains the most commonly used algorithm.

One of the most significant problems with association rules mining is that it often results in too many rules [33], especially when attributes in a data set are highly correlated [22]. On one of our small datasets containing 295 images, several of which had both signal and textual features associated, we extracted over 1.5 million rules when *minsup* and *minconf* were set to 0.02 (or 2%) each and rules were limited to at most two features on the left hand side and one feature on the right hand side. Increasing the Support threshold significantly reduces the number of rules discovered, but risks losing useful associations. In addition, it is non-trivial to set a good value for Support and Confidence thresholds; it depends on the size of data, sparseness of data, and the particular problem under study [6]. Considering these issues, a number of researchers proposed alternate interestingness measures to evaluate and rank discovered associations. These measures are generally divided into subjective and objective interestingness measures. Brijs et al. [6] provides an overview of a number of symmetric objective interestingness measures, five of which are Lift (or Interest), Chi-Square, Correlation Coefficient, Log linear analysis and Empirical Bayes correction. Shekar et al. [28] proposed three measures for capturing relatedness between item pairs. Based on the Chi-Square test, Liu et al. [22] introduces the concept of direction-setting and non-direction-setting rules for summarizing association rules. In a follow up paper [23], they propose a subjective approach that assists the user in finding interesting rules. Hilderman and Hamilton [16] survey various objective and subjective interestingness measures for classification rules, association rules, and generalized relations. Tan et al. [32] discuss the properties of twenty-one objective interestingness measures and analyze the impacts of Support based pruning and contingency table standardization.

2.3 Mining Association Rules from Images

Utilizing object generation capabilities of UC Berkeley's BlobWorld content-based image retrieval system [3, 8], Ordonez and Omiecinski [25] proposed an algorithm to extract association rules from images. The BlobWorld system represents an image as a collection of Blobs. In order to generate association rules, objects extracted by BlobWorld are considered analogous to items and images are considered analogous to transactions. Candidate itemsets are generated from the set of objects, and the Support is calculated by checking individual images for presence or absence of objects. This information is further used to calculate Confidence. This approach works well on a small set of images containing simple geometric objects, but is not suitable for images containing complex objects. Haddad and Mulhem [12] proposed a more realistic approach that considers both manual textual annotations and signal features like dominant colors, directions, and texture indicators to generate association rules from images. Images are first segmented into regions based on their spatial connectivity and visual similarity. Principal color, secondary color, principal direction, and texture features are computed for regions, and annotations are added manually using a list of predefined terms. Finally, association rules are generated using regions as transactions and region features as items. This approach is less scalable and hence not directly applicable to web images.

2.4 Clustering Based on Hypergraph Partitioning

Based on the observation that using association rules directly for clustering would result in clusters that are too granular, Han et al. [14] proposed an approach to cluster transactions using association rule hypergraphs. Hypergraphs are similar to graphs except that each edge, called a hyperedge, could connect two or more vertices. In order to generate a hypergraph from a set of association rules, each unique item that exists in the set is assigned to a unique vertex in the graph. All rules containing a set of items would generate a hyperedge, with average Confidence of such rules used as the weight. For example, if {A} → {B, C} and {C} → {A, B} are all possible rules between items A, B, and C with Confidences 0.6 and 0.4 respectively, there would be a hyperedge between A, B, and C with a weight of 0.5. hMETIS [18], a hypergraph-partitioning algorithm that is widely used in the VLSI domain is used to partition this hypergraph. Transactions are assigned to these partitions using a simple scoring function resulting in clusters of transactions.

hMETIS [18] is a multi-level hyper graph-partitioning algorithm that is based on the multilevel paradigm. In the multilevel paradigm, a sequence of successively coarser hypergraphs is constructed. A bisection of the coarsest hypergraph is computed and it is used to obtain a bisection of the original hypergraph by successively projecting and refining the bisection to the next level finer hypergraph. hMETIS achieves this in three phases. During the coarsening phase, the size of the graph is successively decreased; during the initial partitioning phase, a bisection of the smaller graph is computed; and during the uncoarsening and refinement phase, the bisection is successively refined as it is projected to the larger graphs.

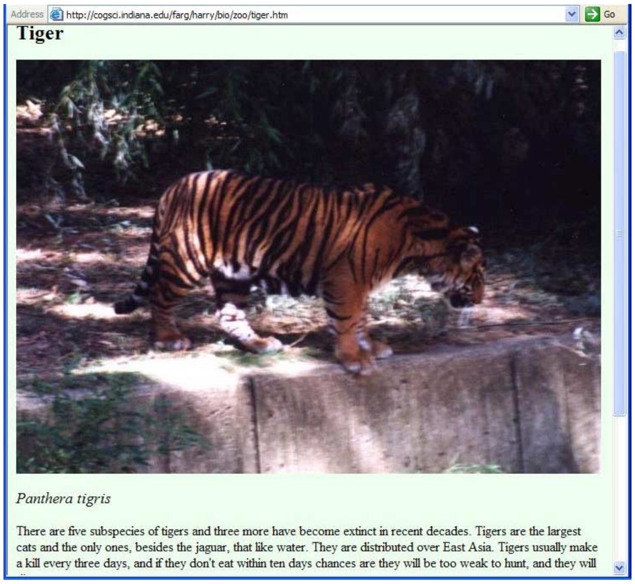
3. MINING ASSOCIATION RULES

3.1 Data Gathering and Preprocessing

Over 3000 images were crawled from the Internet and saved to local disk, along with referring web pages, preserving the links. These images were divided into two separate datasets and the same set of steps was performed on each dataset.

A hashtable was generated using references to images as keys and the lists of their referring web pages as values. All the HTML tags and formatting commands (i.e., "&nbsp;") were stripped out from the web pages, and stopwords were eliminated using the standard list of 571 stopwords initially designed for the SMART system [30]. The remaining terms were stemmed using the Paice stemmer [26]. Since word frequencies within individual documents are believed to be insignificant in the context of web pages [35], they were ignored, and the unique terms from all referring web pages were added as textual features for each image, i.e., the image-word vector was binary valued. Images were scaled to a fixed size of 168 x 168 maintaining their aspect ratios, enabling faster processing times for the signal feature extraction phase.

A web image:



http://cogsci.indiana.edu/farg/harry/bio/zoo/tiger.htm

Tiger

Panthera tigris

There are five subspecies of tigers and three more have become extinct in recent decades. Tigers are the largest cats and the only ones, besides the jaguar, that like water. They are distributed over East Asia. Tigers usually make a kill every three days, and if they don't eat within ten days chances are they will be too weak to hunt, and they will

Textual features extracted from the referring web page(s):

tiger panthera tigris tiger panthera tigris there are five subspecies of tigers and three more have become extinct in recent decades tigers are the largest cats and the only ones besides the jaguar that like water they are distributed over east asia tigers usually make a kill every three days and if they don't eat within ten days chances are they will be too weak to hunt and they will die

Textual features extracted from the image file name:

tiger

Signal features:

@SIGNAL_ORIENTATION=1

@SIGNAL_COLOR=Black

@SIGNAL_COLOR=Brown

@SIGNAL_ORIENTATION=3

Final set of features with textual features stemmed, and duplicates and stopwords eliminated:

@SIGNAL_ORIENTATION=1 hunt die subspecy as @SIGNAL_COLOR=Black distribut kil tig chant extinct wat weak @SIGNAL_COLOR=Brown eat jagu ten east day mak panther tigr cat rec decad largest @SIGNAL_ORIENTATION=3

Figure 1: Features extracted from one of the 3364 images crawled from the web, found at http://cogsci.indiana.edu/farg/harry/bio/zoo/tiger.htm

3.2 Feature Extraction

In addition to the terms extracted from referring web pages, image file names were processed to extract keywords. Terms separated using standard delimiters like space, underscore, and hyphen were isolated and further parsed for potential words, taking case changes and appearance of numbers into account. The resulting keywords were stemmed, checked against the stopwords list, and added to the list of textual features associated with the image.

HSV color histograms were computed and used to identify the two most dominant colors. In order to calculate significant orientations, horizontal and vertical Sobel filters were applied to the image. The resulting values were used to generate a 2D histogram of gradients. Small image gradients were eliminated and the remaining ones were quantized to acquire a coarse representation of the four most significant orientations. The image was then checked for the presence of two major orientations, by comparing the magnitude of the two most significant orientations against the third orientation. If the first two orientations were found to be close to each other but significantly apart from the third orientation, the image usually contained grid like objects; an extra feature indicating this finding was added for such images. The resulting color and orientation features were added as image signal features in a textual form (i.e., Color = BLUE) prefixed as "@SIGNAL_" to avoid potential conflicts with textual features. Color names were assigned to HSV ranges in a way similar to [24], except that we have dealt with relatively fewer colors. Figure 1 shows an image and the set of extracted signal and textual features.

3.3 Rule Generation

In terms of classical association rule terminology, images were considered as transactions, and textual and signal features were considered as items. An algorithm similar to Apriori-TID [2] was used to generate association rules. Large itemsets were computed by checking for the presence or absence of features in images.

Rule	***Support***
{@SIGNAL_COLOR=Brown}→{suv}	0.027118
{@SIGNAL_COLOR=Brown}→{wild}	0.030508
{model}→{car}	0.183050
{jagu}→{turbo}	0.030508
{import}→{hors}	0.027118
{anim}→{@SIGNAL_ORIENTATION=1}	0.054237
{livestock}→ {@SIGNAL_ORIENTATION=1}	0.040677
{@SIGNAL_ORIENTATION=1}→ {@SIGNAL_COLOR=Pink}	0.020338

Figure 2. A few association rules generated from images of cars and animals along with their Support. Note the stemming of "Jaguar", "Horse" and "Animal".

Zipf's law states that items that occur too frequently or very infrequently are not significant, and this has been proven as a useful feature reduction technique in the context of generating associations from web pages [11]. Checking for the Support threshold essentially eliminates infrequent items. We applied an additional feature reduction step on large 1-itemset (i.e., an itemset containing single terms and their frequencies) and eliminated items with very high Support (greater than 0.9). Once generated, rules were written to a file along with their Support, Confidence, and additional information required to calculate the values of various interestingness measures discussed in the next section. Figure 2 shows a few rules extracted from one of our experimental datasets.

4. GENERATING HYPERGRAPHS

A unique vertex was generated from each unique feature that existed in the final set of extracted association rules. A hyperedge was generated between a set of vertices if there was at least one association rule containing exactly the features that existed in the set. As an example, three hyperedges were generated for the following set of four rules:

{Color = YELLOW}→{bart} supp = 0.2, conf = 0.4

{bart}→{Color = YELLOW} supp = 0.2, conf = 0.8

{Color = YELLOW}→{lisa} supp = 0.25, conf = 0.3

{bart} → {lisa} supp = 0.1, conf = 0.5

The first hyperedge was generated between vertices labeled as 'Color = YELLOW' and 'bart', the second hyperedge was generated between vertices labeled as 'Color = YELLOW' and 'lisa', and the third hyperedge was generated between vertices labeled as 'bart' and 'lisa'.

In order to assign weights to these hyperedges, we used one of our set of 22 interestingness measures, taking averages if more than one rule participated in the hyperedge. For example, as in [14], using average "Confidence" of all rules covered by the first hyperedge results in a weight of 0.6 (i.e., the average of 0.4 and 0.8). Similarly, the second and third hyperedges will have 0.3 and 0.5 assigned as weights, respectively.

5. INTERESTINGNESS MEASURES

As discussed above, Support and Confidence are widely criticized as interestingness measures for association rules. For uneven datasets, a high Support threshold results in pruning useful associations between items [32] that are not present in a large number of transactions, and a low Support threshold results in too many rules. Figure 3 shows the number of rules we have obtained

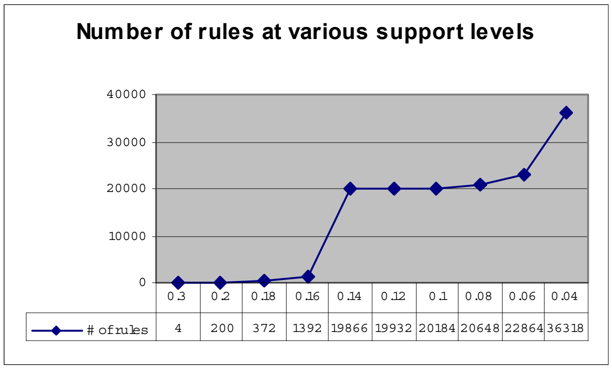


Figure 3. Number of rules generated at various Support levels.

on our smallest dataset of 295 images for various Support levels. Note the non-linearity of the x-axis.

However, [32, 33] show that if the Support threshold is set to a very low value, rules that are pruned contain items that are either uncorrelated or negatively correlated. We use this property to reduce the initial number of rules obtained from our datasets.

On the other hand, Confidence is criticized because of its asymmetric property and its failure to incorporate the baseline frequency of the consequent [4].

Therefore, we experimented using various statistically-inspired interestingness measures as functions to assign weights to hyperedges. Table 1 lists all such measures. Computational details of these measures can be found in [4, 6, 20, 29, 31, 32, 33, 34].

Table 1. List of interestingness measures used

#	*Symbol*	*Interestingness Measure*
1	AV	Added Value
2	F	Certainty Factor
3	χ^2	Chi Square
4	S	Collective Strength
5	c	Confidence
6	V	Conviction
7	Φ	Correlation Coefficient
8	IS	Cosine
9	G	Gini Index
10	I	Interest
11	ζ	Jaccard
12	J	J-Measure
13	κ	Kappa
14	K	Klosgen’s
15	L	Laplace
16	mc	Max Confidence
17	M	Mutual Information
18	α	Odds Ratio
19	RI	Piatetsky Shapiros Interest
20	s	Support
21	Q	Yule's Q
22	Y	Yule's Y

Some of these measures offer properties that can be used to distinguish significant rules from non-significant rules. We used these properties to identify and prune non-significant rules. As an example, Correlation Coefficient and Certainty Factor ranges between –1 and +1 with a value of 0 indicating independence and negative and positive values indicating negative and positive correlation (or in case of certainty factor, dependence) respectively. We used this property to prune all rules having negative or no correlation and kept only the rules containing positively correlated features.

6. CLUSTERING VIA PARTITIONING

In our preliminary experiments, we used a widely used hypergraph partitioning algorithm hMETIS [18] to partition the feature hypergraph. hMETIS is a multi-level partitioning algorithm that has been shown to quickly produce high quality partitions, particularly in its original domain of VLSI design. hMETIS produces balanced k-way partitions where k, the number of partitions, is specified in advance. For this paper, we set the number of partitions based on ground truth.

Once features were partitioned, images were clustered by calculating a score of each image against each partition, based on the features in the partition and the features in the image. All images were assigned to partitions with their highest score. A simple function was used to calculate this score:

$$S = \frac{|I \cap P_i|}{|P_i|}$$

Where I is the set of image features and P_i is the set of features in cluster i.

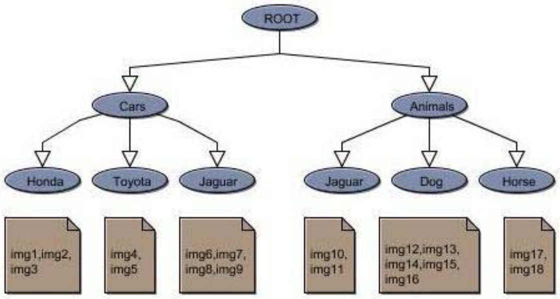


Figure 4. 18 images assigned to ground truth hierarchy.

The main goal of clustering is to organize data in clusters so that intra-cluster similarity is maximized and inter-cluster similarity is minimized [14]. We use a tree-distance-based evaluation measure to evaluate the overall clustering quality, comparing the clustered images with a ground truth hierarchy of image clusters. Each image is evaluated in three ways, and then these individual image scores are summed. The first way assigns a score of ‘p’ to the image for every other image in its cluster that appears in the same ground truth cluster that is the image’s ground truth cluster. The second way deducts a score of ‘n’ from the image for each image in its cluster that appears in a sibling ground truth cluster of the image’s ground truth cluster. The third way deducts a score of ‘z’ for any image in its cluster that does not meet the first two conditions (for example, the image appears in a cousin ground truth cluster rather than a sibling).

As an example, Figure 4 presents a hierarchy with 18 images assigned to various root nodes based on ground truth. If a clustering algorithm generates the following six clusters, we compute the score of cluster 2 as follows:

Cluster1: img5, img9

Cluster2: img2, img10, img1, img8

Cluster3: img3, img12

Cluster4: img14, img18, img17

Cluster5: img11, img16, img13

Cluster6: img4, img6, img7, img15

Using p = z = 1 and n = 0:

img2: (-1 for img10) + (+1 for img1) + (-0 for img8) = 0

img10: (-1 for img2) + (-1 for img1) + (-1 for img8) = -3

img1: (+1 for img2) + (-1 for img10) + (-0 for img8) = 0

img8: (-0 for img2) + (-1 for img10) + (-0 for img1) = -1

Cluster2: 0 – 3 + 0 – 1 = -4.

Scores for other clusters can be calculated in a similar fashion. The overall clustering score is then computed by adding total scores for all clusters. In order to compare clustering quality across datasets of different sizes, max and min bounds for the raw score can be trivially calculated using the ground truth hierarchy, and these extremes can be used to normalize the raw score. We used this technique in this paper; cluster fit therefore is in a range of [0, 1].

Graph partitioning is an NP-hard problem. Efficient partitioning algorithms such as hMETIS [18] use various randomized heuristics to achieve the desired level of performance. A major drawback of this approach is that multiple executions of the algorithm on the same hypergraph using the same parameters often result in different partitions. As suggested in [19], we executed hMETIS a number of times on each feature hypergraph and picked the partition with highest overall clustering score.

7. EXPERIMENTAL RESULTS

A dataset containing 295 images of cars and animals with a ground truth hierarchy as in Figure 4 was used for initial experiments. In the ground truth, the smallest cluster had 7 images and the largest cluster had 80 images. A second dataset containing 3069 images of animals and cartoons was used to validate our results. The smallest cluster in this dataset contained 100 images and the largest cluster contained 1970 images. Both datasets included some categories that could challenge any clustering algorithm because of inherent ambiguity, for example, Jaguar cars and Jaguar animals in the first dataset, and images of ducks and Donald Duck in the second dataset.

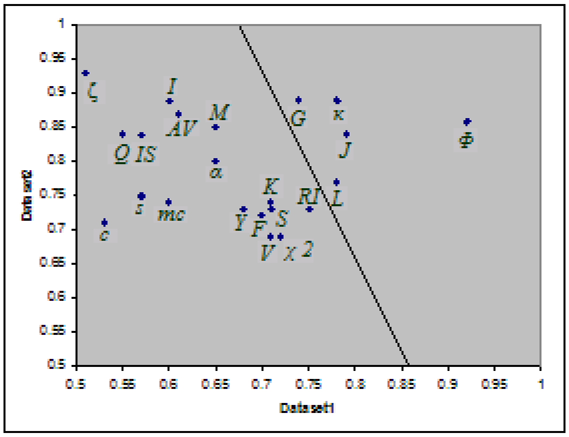
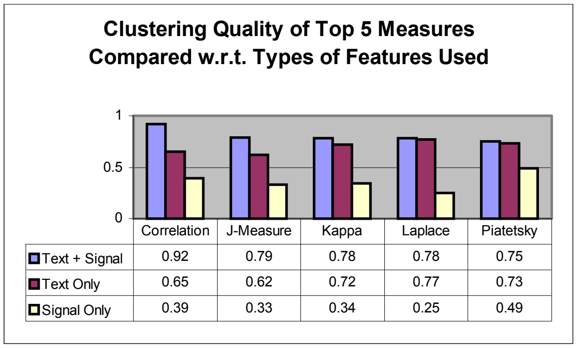


Figure 5. Comparison of clustering quality of various measures across on both datasets.

Additionally, for computational efficiency purposes, these first experiments were performed using rules that contain one item on the left and one item on the right. Figure 5 graphs the overall clustering quality of various interestingness measures on each of these datasets. Clearly, Support and Confidence are among the worst 10 performers on both datasets. Max Confidence, a symmetric version of Confidence, outperformed Confidence on both datasets, which adds credence to the claim that the asymmetric property of Confidence is not as useful in web domain.



	Correlation	J-Measure	Kappa	Laplace	Piatetsky
▨ Text + Signal	0.92	0.79	0.78	0.78	0.75
■ Text Only	0.65	0.62	0.72	0.77	0.73
□ Signal Only	0.39	0.33	0.34	0.25	0.49

Figure 6. Clustering quality comparison on Dataset1 using text-only, signal-only, and both features.

Correlation Coefficient, Kappa, J-Measure, and Gini Index perform consistently well. However Jaccard poses a surprising problem. We suspect that this is due to the imbalance of cluster sizes in the second dataset. It is important to note that the first dataset more clearly determines top 5 measures. Additional experiments were performed on a third, reasonably large, dataset and confirmed the results presented in Figure 5. Details of these additional experiments are omitted for space reasons.

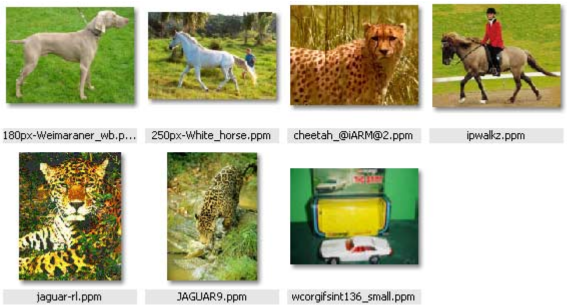


Figure 7. A small cluster generated from the first dataset using signal-only features and Correlation Coefficient as interestingness measure.

Figure 6 compares the clustering quality of the top 5 measures on the first dataset when signal-only, text- only, or both kinds of features are used. Signal-only techniques performed worst in terms of clustering quality. For example, Figure 7 presents a small signal-only cluster generated using Correlation Coefficient on Dataset1. Although all seven images in this cluster look visually similar, they belong to four different semantic categories.

Figure 6 also shows that combining textual and signal features provide improvement over clustering using text-only features. If enough textual features are available, the quality of clustering using text-only features is often comparable to clustering using a combination of textual and signal features. Unfortunately, this is often not the case with web images. Figure 8 presents portion of a cluster generated using text-only features and correlation coefficient on the first dataset. While most of the images may have the keyword 'Jaguar' associated with them, they lacked further information that could have helped separate animals from cars. When signal features were added, the same clustering technique using the same interestingness measure was able to isolate animals and cars in two separate clusters, achieving a much higher level of clustering quality as shown in Figure 9.

7.1 Cross Validation

To validate our findings, leave-n-out cross validation was applied on the first dataset, using two of the top 5 measures as shown in Figure 6 and both signal and textual features used to generate rules. ‘n’ was set to 10, which resulted in 29 unique sets of randomly selected images. 29 experiments were performed for each of the two measures and one of the image sets was left out in each experiment. The remaining images were used to generate rules, and all images from the original dataset (images used to generate rules, as well as images that were left out) were clustered using the hypergraph partitions obtained. Experiments performed using Kappa resulted in an average clustering quality of 0.75, with max = 0.85 and min = 0.70 and experiments performed using J-Measure resulted in an average clustering quality of 0.73 with max = 0.79 and min = 0.69, validating our initial results.

7.2 Rules with More Than Two Features

We performed preliminary experiments to find if hypergraph partitions generated using higher order rules would result in better clustering, as compared to rules that contain only one item on the left and one item on the right. Using a relatively higher support threshold on the first dataset, we generated two sets of rules. The first set contained rules with one item on the left and one item on the right and the second set contained rules with two items on the left and one item on the right. All of the top 5 measures were used to generate hypergraphs that were further used to cluster all images in the dataset. We observed that hypergraphs based on rules containing two items on the left results in an average clustering quality improvement of 19% across all measures, as compared to clustering obtained using hypergraphs containing one item on the left and one item on the right. Specifically, Laplace gained the most and Piatetsky Shapiros Rule Interest gained the least improvement.

8. CONCLUSIONS AND FUTURE WORK

We presented a novel approach of clustering web images using association rules, objective interestingness measures, and hypergraph partitions. We also presented a tree-distance-based clustering evaluation measure that considers the importance of objects ‘occurring together’ based on ground-truth. We showed that statistically-inspired objective interestingness measures, such as Correlation Coefficient and Kappa, result in better clustering as compared to Support and Confidence. Furthermore, we demonstrated that combining textual and signal based features results in better clustering as compared to clustering using signal-only or text-only features. The difference becomes even more prominent when the dataset contains images that belong to different semantic categories but share some textual features.

In the future, we would like to do more detailed experiments using image association rules with more than two features. We also plan to experiment using other hypergraph partitioning algorithms, and to fine-tune our tree-distance-based clustering quality measure.

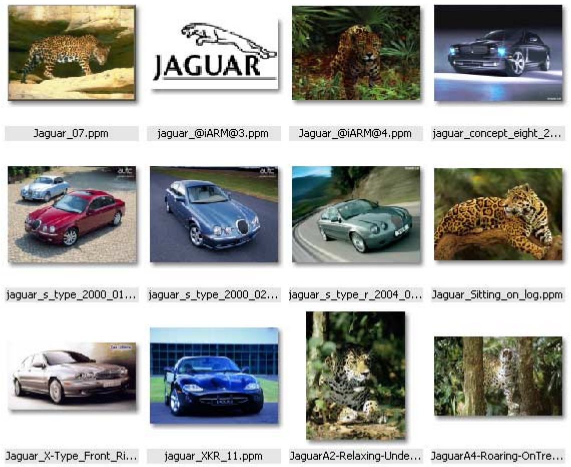

Jaguar_07.ppm
jaguar_@iARM@3.ppm
Jaguar_@iARM@4.ppm
jaguar_concept_eight_2...
jaguar_s_type_2000_01...
jaguar_s_type_2000_02...
jaguar_s_type_r_2004_0...
Jaguar_Sitting_on_log.ppm
Jaguar_X-Type_Front_Ri...
jaguar_XKR_11.ppm
JaguarA2-Relaxing-Unde...
JaguarA4-Roaring-OnTre...

Figure8. Portion of a cluster from the first dataset using text-only features and Correlation Coefficient.

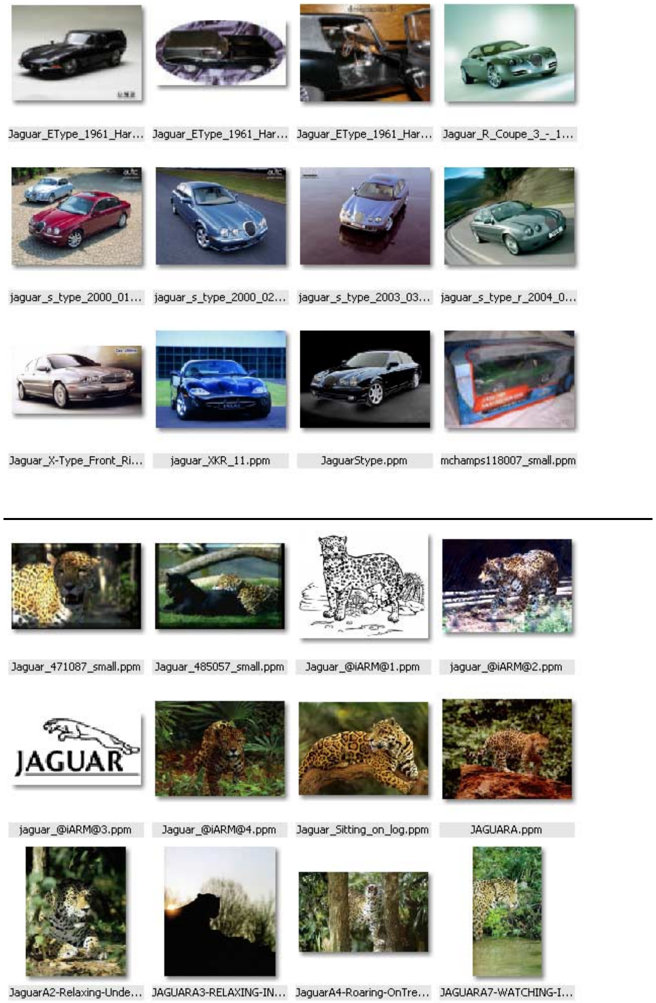


Figure 9. Portions of two clusters from the first dataset using combined textual and signal features and Correlation Coefficient.

9. REFERENCES

[1] Agrawal, R., Imielinski, T. and Swami A. “Mining Association rules between set of items in large databases.” In Proc. *ACM SIGMOD Int. Conf. on Management of Data*, Washington, DC, May 1993.

[2] Agrawal, R. and Srikant, R. “Fast Algorithms for Mining Association Rules in Large Databases.” In Proc. *20th Int. Conf. on Very Large Databases,* pp. 487-499, Santiago, Chile, 1994.

[3] Belongie, S., Carson, C., Greenspan, H. and Malik, J. “Recognition of images in large databases using a learning framework.” *Tech Report TR 97-939, U.C. Berkeley,* 1997.

[4] Berzal, F., Blanco, I., Sánchez, D. and Vila, M.A. "Measuring the Accuracy and Importance of Association Rules: A New Framework". *Intelligent Data Analysis 6:221-235*, 2002.

[5] Bodon, F. “A fast apriori implementation.” In Proc. *IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, 2003.

[6] Brijs**,** T. Vanhoof, K. and Wets, G. “Defining interestingness for association rules.” In *Int. journal of information theories and applications, 10:4,* 2003.

[7] Cai, D., Yu, S., Wen, J.-R. and Ma, W.-Y. “VIPS: a vision based page segmentation algorithm.” *Microsoft Technical Report*, MSR-TR-2003-79.

[8] Carson, C., Belongie, S., Greenspan, H., and Malik, J. “Region-based image querying.” In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1997.

[9] Goethals, B. “Efficient Frequent Pattern Mining.” PhD thesis, Transnational University of Limburg, Belgium, 2002.

[10] Gouda , K. and Zaki, M. J. “Efficiently mining maximal frequent itemsets.” In *1st IEEE Int. Conf. on Data Mining*, Nov. 2001.

[11] Gravano, L., Ipeirotis, P. and Sahami, M. “QProber: A System for Automatic Classification of Hidden-Web Databases.” *In ACM Transactions on Information Systems, vol. 21, no. 1*, Jan. 2003.

[12] Haddad, H., Mulhem, P. **“**Association Rules for Symbolic Indexing of Still Images.” *The 2001 Int. Conf. on Artificial Intelligence*, June, 2001.

[13] Han, E.-H., Karypis, G. and Kumar, V. “Clustering in a high-dimensional space using hypergraph models.” *Tech. Rep. 97-063, University of Minnesota*, 1998.

[14] Han, E.-H., Karypis, G., Kumar, V. and Mobasher, B. “Clustering based on association rule hypergraphs.” In *Research Issues on Data Mining and Knowledge Discovery*, 1997.

[15] He, X., Cai, D., Wen, J.-R., Ma, W.-Y. and Zhang, H.-J. "ImageSeer: Clustering and Searching WWW Images Using Link and Page Layout Analysis." *Microsoft Technical Report, MSR-TR-2004-38*, 2004.

[16] Hilderman R. J., and Hamilton, H. J. "Knowledge Discovery and Interestingness Measures: A Survey." *University of Regina Technical Report, TR 99-04,* 1999.

[17] Hsu, W., Lee, M. L. and Zhang, Ji. “Image Mining: Trends and Developments.” in *Journal of Intelligent Information System (JISS): Special Issue on Multimedia Data Mining*, Kluwer Academic, 2002.

[18] Karypis, G., Aggarwal, R., Kumar, V. and Shekhar, S. “Hypergraph partitioning: Applications in VLSI domain.” *Technical Report TR-96-060, Department of Computer Science, University of Minnesota,* Minneapolis, 1996.

[19] Karypis, G. and Kumar, V. “hMETIS user manual.” *http://www-users.cs.umn.edu/~karypis/metis/hmetis*

[20] Li, J. and Zhang, Y. "Direct Interesting Rule Generation." *icdm, vol. 00, p. 155*, Third 2003.

[21] Lienhart, R. and Hartmann, A. “Classifying images on the web automatically.” *Journal of Electronic Imaging* 11(4), pp. 445-454, Oct 2002.

[22] Liu, B., Hsu, W. and Ma. Y. “Pruning and Summarizing the Discovered Associations.” In Proc. *SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 1999.

[23] Liu, B., Hsu, W., Chen S. and Ma, W.-Y. "Analyzing the Subjective Interestingness of Association Rules." *IEEE Intelligent Systems,* 2000.

[24] Liu, Y., Zhang, D., Lu, G. and Ma, W.-Y. "Region-Based Image Retrieval with High-Level Semantic Color Names," In Proc. *11th Int. Multimedia Modeling Conference,* 2005.

[25] Ordonez, C. and Omiecinski, E. “Discovering Association Rules based on Image Content.” *IEEE Advances in Digital Libraries Conference*, 1999.

[26] Paice, C.D. "Another Stemmer," *SIGIR Forum 24 (3),* 1990.

[27] Rushing, J. A., Ranganath, H. S. and Hinke, T. H. **“**Using Association Rules as Texture Features.” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2001.

[28] Shekar B. and Natarajan, R. "A Transaction-based Neighborhood-driven Approach to Quantifying Interestingness of Association Rules." In Proc. *Fourth IEEE Int. Conf. on Data Mining*, 2004.

[29] Shortliffe, E. and Buchanan, B. "A model of inexact reasoning in medicine." *Mathematical Biosciences 23*, 1975.

[30] SMART Project (eds.) Stopword List for English Information Retrieval, *http://www.unine.ch/info/clef/englishST.txt*

[31] Smyth P. and Goodman, R. M. **“**An information theoretic approach to rule induction from databases.**”** *IEEE Transactions on Knowledge and Data Engineering, Volume 4, Issue 4*, Aug. 1992.

[32] Tan, P., Kumar, V. and Srivastava, J. “Selecting the right interestingness measure for association patterns.” In Proc. *SIGKDD. 32–41*, 2002.

[33] Tan, P. and Kumar, V. “Interestingness measures for association patterns: A perspective,” *KDD 2000 Workshop on Postprocessing in Machine Learning and Data Mining*, 2000.

[34] Viera, A. J. and Garrett, J. M. “Understanding Interobserver Agreement: The Kappa Statistic.” Family Medicine 37(5), p. 360, 2005.

[35] Wong, W.-C. and Fu, A. W.-C. “Incremental Document Clustering for Web Page Classification.” In *2000 Int. Conf. on Information Society in the 21st Century*, 2000.