

**Final Year Project: Final Report**

# **Acquiring structured knowledge from Flickr**

Alex Thomson

BSc Computer Science

1033702

**Project No:** 226

**Supervisor:** Dr. Steven Schockhaert

**Moderator:** Prof. Alun D Preece

## Abstract

The goal of this project is to discover whether location based services can benefit from associative rules derived from a large set of Flickr photo metadata. The rules will reflect common associations between elements (specifically user-defined tags) within the dataset.

The purpose of this final report is to document the implementation, subsequent analysis and conclusion utilising the research acquired from the interim report, more specifically it will cover:

**The initial dataset;** what attributes it contains and how it was manipulated & used.

### Implementations:

- The program used to manipulate the dataset and produce UK samples with all non-essential data removed.
- The program implementing the K-Medoids clustering algorithm to cluster the UK samples and then output each cluster as a suitable transaction list for the Association Rule Mining program Argui.
- The program implementing the Chi-Squared algorithm as a method of feature selection on the tags in the sample to remove undesirable tags.

### Analysis:

- An early evaluation of derived rules from a small sample of photos.
- An evaluation of the derived rules from a much larger sample set and a comparison between results from the differently sized samples.
- After using Chi-Squared to remove undesirable tags from the larger sample there will be a re-evaluation of discovered rules and how effective the feature selection has been.

### Conclusions:

- Whether the steps in this project are efficient and suitable for discovering rules within this dataset.
- A discussion about the types of predictions that have been made from the associations discovered.
- How these rules could be theoretically benefit location based services.

The report concludes with a personal reflection on the project as a whole and many of the changes that could improve the results given the project were restarted.

# Table of Contents

Abstract.....	2
The initial dataset; .....	2
Implementations:.....	2
Analysis: .....	2
Conclusions: .....	2
Glossary.....	5
Acknowledgments.....	5
Introduction .....	6
Project Recap .....	6
Changes since the Interim Report.....	6
Initial Dataset .....	6
Implementation .....	7
Producing a sample subset from the initial dataset .....	7
Program 1- Initial Dataset parsing and UK Sample .....	7
K-Medoids Clustering.....	8
K-Medoids.....	8
The Algorithm .....	9
Program 2 – K Medoids Implementation.....	9
Feature Selection .....	11
Program 3 - Chi-Squared ( $\chi^2$ ) Algorithm and implementation .....	12
Analysis .....	12
Support.....	12
Confidence .....	13
Expectations.....	13
Cluster Sizes .....	13
Early results.....	13
1000 Photos – 4 Clusters.....	14
1000 Photos – 20 Clusters.....	16
Larger Sample Results.....	18
10000 photos - 100 clusters.....	18
10000 photos - 100 clusters including feature selection .....	20
50,000 Photos - 250 Clusters .....	22
Further Considerations .....	23

Sentiment analysis .....	23
Timestamp .....	23
Conclusion .....	23
Implementation .....	24
Analysis .....	24
Holiday Recommendation Service: .....	24
Event Finder Service:.....	24
Tourist Information Service: .....	25
Other Uses: .....	25
Reflection .....	25
Summary .....	25
Preparation .....	25
Implementation .....	26
Personal accomplishments .....	26
Improvements to the implementation .....	26
Analysis .....	27
My approach to analysis .....	27
Improvements to the analysis.....	27
Conclusion.....	28
Improvements to the conclusion .....	28
References .....	30

## Glossary

**Antecedent** – The left hand side of a given rule (Also known as the LHS).

**Argui** – The Association Rule Mining application used for the discovery and analysis of rules.

**ARM** – Association Rule Mining, the process of mining for associations in a given transaction list.

**Chi-Squared ( $\chi^2$ )** – A feature selection method that returns a quantitative value for a given item.

**Consequent** – The right hand side of a given rule (Also known as the RHS).

**Google Maps** – Accessible from [www.maps.google.com](http://www.maps.google.com), it is an online map service.

**K-Medoids Clustering** – The clustering algorithm used in this project to cluster photos by density.

**KML** – Keyhole Mark-up Language, a notation used to write geographical locations that can be interpreted by an online map service.

## Acknowledgments

I would like to express my gratitude to Dr. Steven Schockhaert for his advice on the structure of the project and his valuable insights into the methodology required. It was very much appreciated that I was granted access to his research paper that elaborated on some of methods utilised in this project.

# Introduction

## Project Recap

This project aims to produce meaningful correlations between user defined tags in photos and how these correlations could potentially be utilised by location based services.

The Interim report documented the research into the techniques and methods for deriving associations between elements within datasets. Within the context of this project, that means clustering many photos geographically and searching for correlations between their tags.

The research was covered chronologically relative to how the stages must be executed and this report follows the same pattern.

## Changes since the Interim Report

It was originally decided that it would save time to use an existing K-Medoids clustering implementation, or at least snippets from an existing library; however I felt that my familiarity with the clustering process wasn't sufficient so I decided to implement the algorithm myself. The drawbacks and advantages of this are discussed in its respective section within this report.

## Initial Dataset

The initial dataset was briefly discussed in the interim report, however I shall elaborate on its attributes and how they have been utilised. The dataset itself is a collection of metadata for 16 million photos harvested from Flickr.com and has been made available to me for this project.

The file is approximately 2Gb in size and each line contains 7 comma delimited attributes, these being:

Unique Photo ID	Owner of the Photo	Latitude	Longitude	Tags (themselves separated by a space)	Time Stamp	Flickr Accuracy Level
-----------------	--------------------	----------	-----------	--	------------	-----------------------

**Unique Photo ID** – This is a unique number assigned to each line in the dataset, this attribute was removed at the first stage as it was unnecessary for the clustering algorithm.

**Owner of the Photo** – With users uploading multiple photos, it can be necessary to have an Owner ID, this attribute will be discussed later in the report when considering the issues users that upload many photos.

**Latitude** – Specifies a North-South position of a point on the earth's surface.

**Longitude** - Specifies an East-West position of a point on the earth's surface.

These attributes were used to identify photos within the UK and then again within the clustering algorithm.

**Tags** – User specified words to describe the photo, crucial to the project, these tags will form the basis of rules after the photos have been clustered.

**Time Stamp** – When the photo was taken, this attribute has not been used.

**Flickr Accuracy Level** – This is a numerical value between 1 & 16 that describes how exact the latitude and longitude references are (i.e. A 3 means country, 11 is City, 16 is street), all photos in this set are of maximum accuracy (16) so this attribute is completely unnecessary.

## Implementation

My programming language of choice for this project was Java – it has sufficient functionality to perform what is required and it's what I am personally most familiar with as a programming language.

### Producing a sample subset from the initial dataset

The initial dataset contains 16 million photos from across the world, it was suggested that working with such a large dataset would cause unreasonable computation times when it was time to output clusters for rule mining. It was decided that a small subset of photos from within the UK would be used.

The sample set will first need to be put through the clustering algorithm, so it is essential for the output to be as required for that stage.

To fulfil this necessity I was required to write a program that parses the initial dataset and moves photos from the UK into a sample set, it was also an opportunity to 'clean' the sample up and remove undesired attributes.

### Program 1- Initial Dataset parsing and UK Sample

The first program is relatively lightweight and only performs some simple calculations and reformatting of the dataset for the K-medoids program.

It reads a specified number of lines from the initial dataset and checks to see if its latitude and longitude fall within the outermost boundaries of the UK, that is:

A Latitude value between the northernmost point 58.66 and the southernmost point 49.85

A Longitude value between the easternmost point -13.68 and the westernmost point 1.76

For each line that is within the UK, it is stripped of its unnecessary attributes (Unique ID, Owner ID, Time Stamp, and Flickr Accuracy Level) and output to the next line on the sample text file containing its latitude, longitude and tags.

This program can be found in appendix A.

## K-Medoids Clustering

During the research for this project, I believe I underestimated the importance of this algorithm and how necessary it was to understand it thoroughly. It is vital that I have the capability of easily creating multiple samples with differing cluster sizes to compare and analyse correlations and rules.

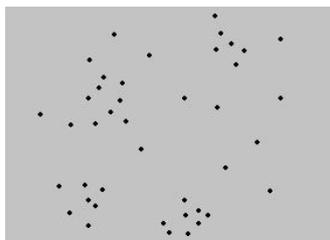
I had originally planned to search for an existing implementation (there was one available (Abeel, 2006)) to save time and effort unnecessarily re-implementing this clustering algorithm, however after realising the importance of its functionality I decided to implement it personally and although it was fairly time consuming, I feel I have a much superior understanding of how it works and as a further benefit, I was able to tailor the implementation around the datasets I'm using.

### K-Medoids

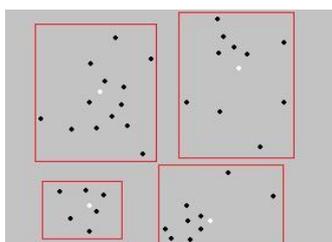
K (Mirkes, 2011) a set of datapoints along with a specified amount of desired clusters and produce a set of density driven clusters.

Within the context of this project, that means taking a set of photos (each with coordinates) and producing a desired number of clusters that each contains photos (with their tags) within close geographical proximity of each other.

For example, if faced with a potential set of datapoints like:



The algorithm would be expected to produce something similar to this if it was asked for 4 clusters:



## The Algorithm

The algorithm only requires a dataset (of datapoints, each with coordinates) and X number of clusters as input:

1. Firstly select X random datapoints from the dataset to be initial medoids.
2. Assign each datapoint in the dataset to its closest Medoid using whichever distance method is more appropriate (I used Euclidean Distance Measuring, however if the sample size was any larger, great circle distance measuring would have been more appropriate given the spherical nature of the planet).
3. You now have your clusters and each cluster has a Medoid.
4. Now recalculate the Medoid for each cluster by going through each datapoint in the cluster and seeing which one has the shortest combined distance to every other datapoint.
5. Now with a new set of medoids, you reassign the whole dataset again to their closest Medoid, resulting in the new set of clusters.
6. You repeat steps 4 and 5 several times to allow the Medoids to centralise and settle.

(Mirkes, 2011)

## Program 2 – K Medoids Implementation

My implementation has 2 configurable variables to alter the output as desired:

- How many clusters are required (X)
- How many times do the medoids need recalculating (Y)

Following the algorithm, my program reads the sample dataset and performs the following:

1. Read in the sample dataset and create an array of photos.
2. Randomly pick X (non-repeating) photos to be the initial medoids.
3. Each photo is assigned to its closest Medoid (using Euclidean Distance Measuring with the coordinates). All photos with the same Medoid are essentially a cluster.
4. In every cluster and for each photo, the distance between itself and every other photo (in that cluster) is calculated and accumulated, the photo with the lowest combined distance is there for the most central geographically and becomes the new Medoid for that cluster.
5. Each photo from the sample set (disregarding existing clusters) is reassigned to its closest Medoid (potentially the same or different) and new clusters are formed.
6. Steps 5 and 6 are repeated Y times.

Keyhole Mark-up Language (KML) is an xml notation for expressing geographical locations (Google, 2008) on 2d/3d earth browsers. Google Maps has the ability to interpret KML files and display the data onto a 2d map.

This means I can output the Latitude and Longitude coordinates of any photo (or photos) and store them in a KML file, upload the file online and use Google Maps to display the data. Since writing a program to take coordinates and change them into KML notation would be rather wasteful of time, I have decided to use an already existing macro (Simon, 2007) that takes Latitude and Longitude coordinates and outputs them to a KML file.

I have output the first 1000 UK photos from the dataset and converted their coordinates to KML and uploaded them online, feeding that file link into Google Maps has returned -



I have then input the 1000 photo sample into the K-Medoids implementation with 4 clusters to be output, I have then used the macro to produce a KML file for each cluster, outputting each one separately to Google Maps to demonstrate each cluster visually.



It is quite evident the sample set has been clustered into regions, however a larger sample set and many more clusters are required to hone in on specific pockets of density, and this is where associations should be most valuable.

My implementation of this Algorithm can be seen in appendix B; however this implementation also includes several extra pieces of functionality; such as formatting the clusters so they can then be input directly into the Argui.

## Feature Selection

It is important to note that implementing feature selection came towards the end of this project, however since it required implementing the Chi Squared algorithm, I have included it in this section to help keep the report better organised.

Feature selection is the process of deciding whether and what attributes or values are desirable, and if they aren't, removing them from dataset. Within this project, that means looking at tags within clusters and deciding how meaningful or useful they are, and removing them from the consideration of associations if they aren't.

The method selected during the research for this project was the Chi Squared ( $\chi^2$ ) algorithm, Chi Squared can be used to evaluate a set of qualitative data and return a quantitative value to which other elements in the set have relative values (and can then be compared).

For the purpose of this project, that means identifying how meaningful or useful an individual tag may be within a cluster. This is done by measuring the frequency of a tags appearance within a cluster relative to the frequency of which it appears outside of the cluster (relative to every other tags frequency in and outside the cluster). The end result is a list (for every cluster) of every tag that occurs and their corresponding value that this algorithm has provided, from here all tags below a certain threshold can be removed and hopefully more meaningful associations can be derived from the remaining tags.

An example of this is during the early results, when a small amount of clusters returned a similar top rule where 'England' and 'UK' were found commonly associated with. As the sample is only taken from the UK these discoveries are reasonable yet almost completely useless, any associations these tags are part of are obvious - they exist in a UK subset of photos! 'England' is slightly more reasonable considering it won't be as prevalent in photos taken in Scotland, Wales and Ireland. As we are looking for rules that are specific to the clusters, it is not ideal to have tags that occur frequently across many clusters, instead the more meaningful and specific rules will be found by looking at tags that occur many times in fewer clusters.

## Program 3 - Chi-Squared ( $\chi^2$ ) Algorithm and implementation

$$\chi^2(c, t) = \frac{(O_{tc} - E_{tc})^2}{E_{tc}} + \frac{(O_{t\bar{c}} - E_{t\bar{c}})^2}{E_{t\bar{c}}} + \frac{(O_{\bar{t}c} - E_{\bar{t}c})^2}{E_{\bar{t}c}} + \frac{(O_{\bar{t}\bar{c}} - E_{\bar{t}\bar{c}})^2}{E_{\bar{t}\bar{c}}} \quad (\text{O, S, \& B, 2012})$$

The algorithm calculates a value for a given tag  $t$  in a given cluster  $c$  where:

- $O_{tc}$  is the number of photos in  $c$  that  $t$  occurs
- $O_{t\bar{c}}$  is the number of photos outside  $c$  that  $t$  occurs (i.e. all other clusters)
- $O_{\bar{t}c}$  is the number of photos in  $c$  that  $t$  doesn't occur
- $O_{\bar{t}\bar{c}}$  is the number of photos outside  $c$  that  $t$  doesn't occur

$E_{tc}$  is the expected occurrences of  $t$  in  $c$  and is defined as  $N * P(t) * P(c)$  where:

- $N$  is the total number of photos in the sample
- $P(t)$  is the probability of a tag occurring in any photo, calculated by dividing the total occurrences of  $t$  by the occurrences of all tags
- $P(c)$  is the probability that a photo is located in  $c$ , calculated by dividing the amount of photos in  $c$  by the total number of photos
- Therefore  $E_{tc} = N * P(t) * P(c)$ ,  $E_{t\bar{c}} = N * P(t) * (1 - P(c))$ ,  $E_{\bar{t}c} = N * (1 - P(t)) * P(c)$ ,  $E_{\bar{t}\bar{c}} = N * (1 - P(t)) * (1 - P(c))$

Again this was implemented using Java and because it iterates over the unique tags in each cluster, I could simply program it to read the same input that the rule mining program Argui accepts, however this can iterate over all of the cluster files and output values at the end for each.

The input is read into an array of clusters, each holding an array of photos, in turn each holding an array of tags. A unique vocabulary of tags is created for each cluster and each tags value is calculated using the formula outlined above. The implementation for this can be found in appendix C

## Analysis

The output from the K-Medoids implementation is a set of text files (clusters) that contain rows of tags, where each line represents a different photo.

This is the transaction list that Association Rule Mining (discussed in the interim report) uses to derive correlations between elements or sets of elements (tags in this case).

A rule is essentially a strong correlation and requires an antecedent and a consequent; i.e. both sides of the rule (If  $X$  occurs then  $Y$  occurs). When deriving rules from such a large dataset, there are two important figures that help narrow the types of rules discovered:

**Support:** Support is essentially how frequent the first side of the rule (antecedent) appears within the sample compared to the size of the sample. For example we may find an association that only happens once within a million transactions, however because the one time that particular association occurs it is true, it has maximum confidence and will be discovered as a top rule (without a support threshold). It is important to have a reasonable amount of support, it will likely be a case of manually lowering the threshold until more and more rules are presented.

**Confidence:** This is a ratio of how often the antecedent and consequent occur together against how often the antecedent occurs without the consequent. A confidence of 100 means that every time the consequent occurs, the antecedent does as well; this is potentially a very valuable rule (providing it has enough support).

## Expectations

Regarding the results, there are some general expectations about the types of associations that may be discovered.

**Cluster Sizes** – cluster sizes and how far they span geographically will depend on the amount of clusters the sample is clustered in to. The larger the span geographically, the broader the tags should be (the place names) which should mean that the highest ranked associations could be between place names that correspond to that cluster.

If the amount of clusters is increased, the clusters should start to become dense pockets of photos scattered around the UK, a denser area of photos (For example London) should mean more clusters.

I still expect associations to be mainly focused on names of locations as the amount of clusters increase; however they will become more specific.

After the feature selection method Chi Squared has been applied to the clusters, it should return a list of tags within each cluster ranked by a value that depicts their usefulness, it should at this point have lowered the importance of some of the broader geographical tags (England, UK etc.) and increased the importance of cluster specific tags that occur frequently (an example could be 'Millennium Stadium' in a cluster that has photos from Cardiff).

Once the clusters have been reiterated over and the less useful tags have been removed, association should then appear to be much more region specific and have potential benefits.

The most valuable types of associations that might be discovered are links between the geographical location of the cluster, and something specific to that region, but not a place or a name, rather an event, occasion or attribute.

## Early results

To establish a better understanding of the types of rules that can be discovered, I chose to start with a small sample set of 1000 photos, using my K-medoids implementation to cluster them in to 4 clusters. It was at this stage I noticed an important oversight during my research stage; the association rule mining application Argui doesn't have the capability of reading in multiple files (i.e. clusters) and instead requires the user to manually load a file and then output the rules.

Although this isn't necessarily an issue when looking at four clusters, I will be testing cluster sizes of 10/100/1000 on increasingly larger sample sizes – manually examining all of them isn't feasible as it would require loading each file and manually assessing the rules – potentially hundreds of hours of examination.

As the project is in its latter stages, it would have been too time consuming to find another suitable implementation (None of the other applications from the research stage were suitable); it would have also been complicated to manipulate the programs I have built as they are all designed to output files specifically for Argui.

To keep the solution simple, I decided to continue to use Argui but limit the amount of clusters I analyse from each set to small number, starting with the densest cluster.

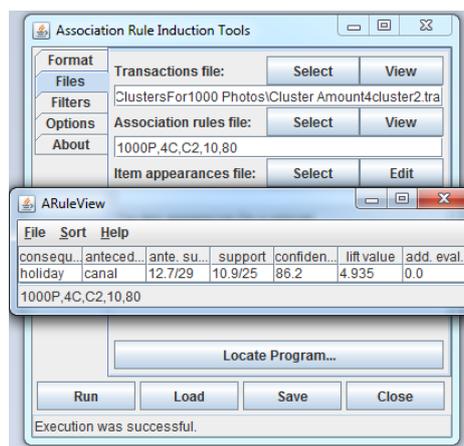
### 1000 Photos – 4 Clusters

It's important to note that since the clustering implementation is written in java and the output is an array of clusters, the index starts at 0.

- Cluster 0 has 140 photos
- Cluster 1 has 125 photos
- Cluster 2 has 543 photos
- Cluster 3 has 192 photos

### 1000 Photos – 4 Clusters - Cluster 2 – 543 photos

Argui has a default setting of 10% minimum support (for the antecedent) and 80% minimum confidence (for the association) and originally found 1 rule given these conditions:



Rather the lower the confidence level to see if more rules would be discovered, I decided lowering the support was probably more conducive to discovering more hidden rules that still had enough confidence to later make them 'predictions'.

Lowering the support to 5% revealed 19 rules, Argui has a built in feature to view the rule set:

consequent	antecedent	ante. supp.	support	confidence	lift value	add. eval.
holiday	bridge	5.2/12	4.4/10	83.3	4.771	0.0
canal	bridge	5.2/12	4.4/10	83.3	6.58	0.0
holiday	avonring	7.4/17	7.4/17	100.0	5.725	0.0
canal	avonring	7.4/17	6.1/14	82.4	6.503	0.0
holiday	rochdalecanal	9.2/21	9.2/21	100.0	5.725	0.0
hisforhome	vintage	8.3/19	8.3/19	100.0	10.409	0.0
vintage	hisforhome	9.6/22	8.3/19	86.4	10.409	0.0
nikon	d50	5.7/13	5.7/13	100.0	15.267	0.0
d50	nikon	6.6/15	5.7/13	86.7	15.267	0.0
geotagged	d50	5.7/13	5.7/13	100.0	10.409	0.0
holiday	canal	12.7/29	10.9/25	86.2	4.935	0.0
geotagged	nikon	6.6/15	5.7/13	86.7	9.021	0.0
canal	avonring holiday	7.4/17	6.1/14	82.4	6.503	0.0
holiday	avonring canal	6.1/14	6.1/14	100.0	5.725	0.0
geotagged	d50 nikon	5.7/13	5.7/13	100.0	10.409	0.0
nikon	d50 geotagged	5.7/13	5.7/13	100.0	15.267	0.0
d50	nikon geotagged	5.7/13	5.7/13	100.0	17.615	0.0

We can see from these results that there are only a few potentially meaningful associations, the appearance of 'holiday' is perhaps indicative that whichever geographical location this cluster covers may be considered as a holiday destination, however given the sample set is quite small and the amount of clusters is only four, it's likely this cluster covers a relatively vast region.

Another tag that might be useful is 'canal' which although associated in this case with 'bridge' was obviously supported enough (i.e. the tags were found in enough photos) possibly suggesting a famous canal or popular attraction of this geographical location. Although it was only possible to spot it manually, canal also appears in the 5th rule as part of 'rochdalecanal' associated with 'holiday', an online search for Rochdale Canal shows it as a popular canal circuit part of the Avon Ring canal ring (Also mentioned rules 3 & 4). Again this further reinforces the likelihood that photos were taken at the location as part of a holiday, possibly making this useful for holiday recommendation.

Although the tag 'holiday' is particularly useful within the goals of this project, I worry that it's likely frequency across many clusters will mean it is removed during feature selection, which values items that occur many times in fewer clusters.

The rest of the tags were slightly more self-descriptive; for example 'geotagged', 'd50 nikon', 'nikon' are all references to the actual taking and documentation of the photo (d50 nikon is a camera and geotagged is a process of tagging photos with a geographical location). Tags like this are less likely to be picked up through feature selection because they are more likely a result of a user's tagging habits – someone who has uploaded many photos of an object/event/location (they all appear in the same place geographically) and tagged every one with the same tags, this will cause problems because it's possible this user may only have photos in one area and therefore these highly prevalent tags in a singular cluster are exactly what this project desires, and what feature selection will pick up as high value tags.

Sets of discovered rules will always need to be manually checked and analysed regardless, however reducing the 'noise' of these meaningless tags could be done a few ways:

1. User ID – although this attribute was removed early on in my implementation because it seemed unnecessary, it would be possible to potentially use this count many occurrences of a tag from a single user as 1 occurrence. Thus stopping what is essentially user defined associations.
2. Build a list of undesirable words; such as generic terms (England, UK, Sunny) or self-describing terms (“Photo”, “Camera” or “Tagged” etc.) and have these tags removed from the sample set.

If it was ever decided that tags such as ‘holiday’ weren’t desirable, possibly because it was knowledge that the cluster we’re looking at is a holiday resort and the discovery of more specific association is more important, it would be possible to have a maximum support – that is to say, items that have high or very high support within a cluster could be automatically discarded.

### 1000 Photos – 20 Clusters

- Cluster 0 has 80 photos
- Cluster 1 has 4 photos
- Cluster 2 has 43 photos
- Cluster 3 has 40 photos
- Cluster 4 has 38 photos
- Cluster 5 has 34 photos
- Cluster 6 has 28 photos
- Cluster 7 has 89 photos
- Cluster 8 has 166 photos
- Cluster 9 has 27 photos
- Cluster 10 has 21 photos
- Cluster 11 has 97 photos
- Cluster 12 has 52 photos
- Cluster 13 has 54 photos
- Cluster 14 has 61 photos
- Cluster 15 has 5 photos
- Cluster 16 has 17 photos
- Cluster 17 has 86 photos
- Cluster 18 has 8 photos
- Cluster 19 has 50 photos

### 1000 Photos – 20 Clusters - Cluster 8 – 166 photos

Starting with the same figures, more rules were discovered, I suspect this is because the support measures a tags occurrences relative to the size of the amount of tags in the cluster.

consequent	antecedent	ante_supp	support	confidence	lift value	add. eval.
london		100.0/166	80.1/133	80.1	1.0	0.0
nikon	d200	12.0/20	12.0/20	100.0	7.905	0.0
d200	nikon	12.7/21	12.0/20	95.2	7.905	0.0
england	d200	12.0/20	12.0/20	100.0	2.912	0.0
london	d200	12.0/20	12.0/20	100.0	1.248	0.0
england	nikon	12.7/21	12.7/21	100.0	2.912	0.0
london	nikon	12.7/21	12.7/21	100.0	1.248	0.0
england	unitedkingdom	13.3/22	12.0/20	90.9	2.648	0.0
london	unitedkingdom	13.3/22	13.3/22	100.0	1.248	0.0
london	uk	15.1/25	15.1/25	100.0	1.248	0.0
london	england	34.3/57	34.3/57	100.0	1.248	0.0
england	d200 nikon	12.0/20	12.0/20	100.0	2.912	0.0
nikon	d200 england	12.0/20	12.0/20	100.0	7.905	0.0
d200	nikon england	12.7/21	12.0/20	95.2	7.905	0.0
london	d200 nikon	12.0/20	12.0/20	100.0	1.248	0.0
nikon	d200 london	12.0/20	12.0/20	100.0	7.905	0.0
d200	nikon london	12.7/21	12.0/20	95.2	7.905	0.0
london	d200 england	12.0/20	12.0/20	100.0	1.248	0.0
england	d200 london	12.0/20	12.0/20	100.0	2.912	0.0
london	nikon england	12.7/21	12.7/21	100.0	1.248	0.0
england	nikon london	12.7/21	12.7/21	100.0	2.912	0.0
london	unitedkingdom england	12.0/20	12.0/20	100.0	1.248	0.0
england	unitedkingdom london	13.3/22	12.0/20	90.9	2.648	0.0
london	d200 nikon england	12.0/20	12.0/20	100.0	1.248	0.0
england	d200 nikon london	12.0/20	12.0/20	100.0	2.912	0.0
nikon	d200 england london	12.0/20	12.0/20	100.0	7.905	0.0
d200	nikon england london	12.7/21	12.0/20	95.2	7.905	0.0

The types of rules discovered are fairly similar to before; 'London' is perhaps the most useful tag being associated with various other tags as it gives an indication of location for the cluster.

The remaining associations are mostly combinations of useless tags. As these conditions returned 27 rules, I have decided to try and find a way of eliminating these highly prevalent and meaningless rules. Firstly, since I know that these rules are discovered because their support is between 10% and 100%, if I change the values to show discoveries between 5% and 10% it will at least display a new set of rules (albeit lower in support).

consequent	antecedent	ante. supp.	support	confidence	lift value	add. eval.
london	londoneye	6.0/10	6.0/10	100.0	1.248	0.0
london	westminster	6.6/11	6.6/11	100.0	1.248	0.0
london	street	6.0/10	6.0/10	100.0	1.248	0.0
london	geotagged	6.0/10	5.4/9	90.0	1.123	0.0
nikon	d200 unitedkingdom	7.2/12	7.2/12	100.0	7.905	0.0
d200	nikon unitedkingdom	7.2/12	7.2/12	100.0	8.3	0.0
england	d200 unitedkingdom	7.2/12	7.2/12	100.0	2.912	0.0
london	d200 unitedkingdom	7.2/12	7.2/12	100.0	1.248	0.0
england	nikon unitedkingdom	7.2/12	7.2/12	100.0	2.912	0.0
london	nikon unitedkingdom	7.2/12	7.2/12	100.0	1.248	0.0
london	uk england	9.6/16	9.6/16	100.0	1.248	0.0
england	d200 nikon unitedkingdom	7.2/12	7.2/12	100.0	2.912	0.0
nikon	d200 unitedkingdom engl...	7.2/12	7.2/12	100.0	7.905	0.0
d200	nikon unitedkingdom engl...	7.2/12	7.2/12	100.0	8.3	0.0
london	d200 nikon unitedkingdom	7.2/12	7.2/12	100.0	1.248	0.0
nikon	d200 unitedkingdom london	7.2/12	7.2/12	100.0	7.905	0.0
d200	nikon unitedkingdom lond...	7.2/12	7.2/12	100.0	8.3	0.0
london	d200 unitedkingdom engl...	7.2/12	7.2/12	100.0	1.248	0.0
england	d200 unitedkingdom london	7.2/12	7.2/12	100.0	2.912	0.0
london	nikon unitedkingdom engl...	7.2/12	7.2/12	100.0	1.248	0.0
england	nikon unitedkingdom lond...	7.2/12	7.2/12	100.0	2.912	0.0
london	d200 nikon unitedkingdo...	7.2/12	7.2/12	100.0	1.248	0.0
england	d200 nikon unitedkingdo...	7.2/12	7.2/12	100.0	2.912	0.0
nikon	d200 unitedkingdom engl...	7.2/12	7.2/12	100.0	7.905	0.0
d200	nikon unitedkingdom engl...	7.2/12	7.2/12	100.0	8.3	0.0

Although there are still countless occurrences of the aforementioned useless tags, there are some interesting newer rules near the top – London & LondonEye, London & Westminster are both much more area specific with the LondonEye being an iconic Ferris wheel within London and Westminster being a place within London.

If I again change the support to show associations between 4% and 5% I get 75 more rules, with only one seemingly interesting link between London & Museum.

Between 3% and 4% I get 352 rules, still with a majority of noise but the rules are slightly more specific again as links between London & UKForeignOffice, London & Architecture are discovered.

Between 2% and 3% are only 75 rules; however the amount of undesirable rules has dropped dramatically:

consequent	antecedent	ante. supp.	support	confidence	lift value	add. eval.
tartalom	christophersweeney	2.4/4	2.4/4	100.0	41.5	0.0
christophersweeney	tartalom	2.4/4	2.4/4	100.0	41.5	0.0
museum	british	2.4/4	2.4/4	100.0	23.714	0.0
london	british	2.4/4	2.4/4	100.0	1.248	0.0
d200	regentspark	2.4/4	2.4/4	100.0	8.3	0.0
nikon	regentspark	2.4/4	2.4/4	100.0	7.905	0.0
england	regentspark	2.4/4	2.4/4	100.0	2.912	0.0
london	regentspark	2.4/4	2.4/4	100.0	1.248	0.0
london	bigben	2.4/4	2.4/4	100.0	1.248	0.0
london	leicestersquare	2.4/4	2.4/4	100.0	1.248	0.0
london	southbank	2.4/4	2.4/4	100.0	1.248	0.0
square	squareformat	2.4/4	2.4/4	100.0	33.2	0.0
squareformat	square	3.0/5	2.4/4	80.0	33.2	0.0
london	sw1	2.4/4	2.4/4	100.0	1.248	0.0
london	building	2.4/4	2.4/4	100.0	1.248	0.0
london	regentstreet	2.4/4	2.4/4	100.0	1.248	0.0
geotagged	canon40d	2.4/4	2.4/4	100.0	16.6	0.0
england	britain	3.0/5	2.4/4	80.0	2.33	0.0
london	1755mm	3.0/5	2.4/4	80.0	0.998	0.0
geotagged	road	3.0/5	2.4/4	80.0	13.28	0.0
england	city	3.0/5	2.4/4	80.0	2.33	0.0
london	british museum	2.4/4	2.4/4	100.0	1.248	0.0
museum	british london	2.4/4	2.4/4	100.0	23.714	0.0
nikon	regentspark d200	2.4/4	2.4/4	100.0	7.905	0.0
d200	regentspark nikon	2.4/4	2.4/4	100.0	8.3	0.0

What this is suggesting is that discovering meaningful rules for any given cluster will likely mean altering the support in small variations until suitable amount and type of rules are discovered.

## Larger Sample Results

### 10000 photos - 100 clusters

As an intermediate step before calculating clusters and associations for a large sample size (there is a significant runtime increase as the dataset increases), I will briefly look at how similar the rules are derived with a dataset of this size. The largest cluster of the 100 created was cluster 40.

consequent	antecedent	ante. supp.	support	confidence	lift value	add. eval.
ireland		100.0/267	82.0/219	82.0	1.0	0.0
dublin		100.0/267	91.4/244	91.4	1.0	0.0
ireland	party	10.9/29	10.5/28	96.6	1.177	0.0
dublin	party	10.9/29	10.5/28	96.6	1.057	0.0
dublin	2008	11.2/30	10.5/28	93.3	1.021	0.0
ireland	drinks	13.9/37	13.9/37	100.0	1.219	0.0
dublin	drinks	13.9/37	13.9/37	100.0	1.094	0.0
tkd	ucd	11.6/31	10.5/28	90.3	6.89	0.0
ucd	tkd	13.1/35	10.5/28	80.0	6.89	0.0
club	ucd	11.6/31	11.2/30	96.8	6.983	0.0
ucd	club	13.9/37	11.2/30	81.1	6.983	0.0
ireland	ucd	11.6/31	11.6/31	100.0	1.219	0.0
dublin	ucd	11.6/31	11.6/31	100.0	1.094	0.0
ireland	night	12.0/32	11.2/30	93.7	1.143	0.0
dublin	night	12.0/32	12.0/32	100.0	1.094	0.0
club	tkd	13.1/35	13.1/35	100.0	7.216	0.0
tkd	club	13.9/37	13.1/35	94.6	7.216	0.0
ireland	tkd	13.1/35	13.1/35	100.0	1.219	0.0
dublin	tkd	13.1/35	13.1/35	100.0	1.094	0.0
ireland	club	13.9/37	13.9/37	100.0	1.219	0.0
dublin	club	13.9/37	13.9/37	100.0	1.094	0.0
dublin	ireland	82.0/219	82.0/219	82.0	1.074	0.0

It becomes quickly apparent that there are many more desirable tags and less self-descriptive tags, I suggest this is because as the amount of photos in a cluster increases,

support for popular descriptive terms used by many users outweigh terms that are only found in one users photoset.

Looking at the associations themselves, it seems obvious the cluster is geographical centred somewhere in Dublin in Ireland, the discovery between 'Dublin' and 'party' has a potential for use but it seems like such little meaningful reward for a comparatively large amount of manual analysis. After lowering the support values to between 5% and 10% associations between month names began to appear.

It does seem reasonable that users may tag their photos with a date or month, these terms on their own or simply associated with a generic term aren't particularly revealing, for example:

consequent	antecedent	ante. supp.	support	confidence	lift value	add. eval.
ireland	august	6.0/16	6.0/16	100.0	1.219	0.0

August => Ireland isn't a particularly revealing association, however if rules were discovered where the consequent was an itemset of more than one tag, date/month/period identifiers could be potentially very helpful, for example if August => Ireland,Party was a very high confidence/support association, it might show August as a popular time to travel to Ireland for a party.

I made sure to browse the remaining rules for associations of this kind but none were apparent, a further 93 rules were found between 4-5% and then 352 between 3-4%, there were some slightly more relevant discoveries as the support was dropped:

consequent	antecedent	ante. supp.	support	confidence	lift value	add. eval.
ireland	christmas dublin	3.4/9	3.4/9	100.0	1.219	0.0
dublin	october ireland	3.4/9	3.4/9	100.0	1.094	0.0
ireland	october dublin	3.7/10	3.4/9	90.0	1.097	0.0
club	karaoke tkd	3.4/9	3.4/9	100.0	7.216	0.0
tkd	karaoke club	3.4/9	3.4/9	100.0	7.629	0.0
ireland	karaoke tkd	3.4/9	3.4/9	100.0	1.219	0.0
tkd	karaoke ireland	4.1/11	3.4/9	81.8	6.242	0.0
dublin	karaoke tkd	3.4/9	3.4/9	100.0	1.094	0.0
tkd	karaoke dublin	4.1/11	3.4/9	81.8	6.242	0.0
ireland	karaoke club	3.4/9	3.4/9	100.0	1.219	0.0
club	karaoke ireland	4.1/11	3.4/9	81.8	5.904	0.0
dublin	karaoke club	3.4/9	3.4/9	100.0	1.094	0.0
club	karaoke dublin	4.1/11	3.4/9	81.8	5.904	0.0
ireland	messrsmaguires party	3.4/9	3.4/9	100.0	1.219	0.0
dublin	messrsmaguires party	3.4/9	3.4/9	100.0	1.094	0.0
club	messrsmaguires ucd	3.7/10	3.7/10	100.0	7.216	0.0
ucd	messrsmaguires club	3.7/10	3.7/10	100.0	8.613	0.0
ireland	messrsmaguires ucd	3.7/10	3.7/10	100.0	1.219	0.0
dublin	messrsmaguires ucd	3.7/10	3.7/10	100.0	1.094	0.0
ireland	messrsmaguires club	3.7/10	3.7/10	100.0	1.219	0.0
dublin	messrsmaguires club	3.7/10	3.7/10	100.0	1.094	0.0
ireland	leaving party	3.7/10	3.4/9	90.0	1.097	0.0
dublin	leaving party	3.7/10	3.4/9	90.0	0.985	0.0
ireland	leaving drinks	3.4/9	3.4/9	100.0	1.219	0.0
dublin	leaving drinks	3.4/9	3.4/9	100.0	1.094	0.0
ireland	birthday party	3.7/10	3.7/10	100.0	1.219	0.0
dublin	birthday party	3.7/10	3.7/10	100.0	1.094	0.0
tkd	2010 ucd	3.4/9	3.4/9	100.0	7.629	0.0
ucd	2010 tkd	3.7/10	3.4/9	90.0	7.752	0.0
club	2010 ucd	3.4/9	3.4/9	100.0	7.216	0.0

Event tags like karaoke, party and birthday are all useful identifiers given they are paired with a specific location. To get a visual depiction of how specific this cluster is, I have plotted it Google Maps.



This cluster consists of 267 photos in the heart of Dublin, Ireland. It's from knowledge about the location of the clusters themselves that can assist in bringing further meaning to discovered associations. By this I mean if two location independent items are found associated, for example Birds=>Rare, it is obviously imperative to know the location of the cluster that contains this association because once known, the rule can be expressed more suitably for a travel recommendation engine as Location => Rule.

### 10000 photos - 100 clusters including feature selection

The chosen feature selection method for this project is the Chi Squared algorithm, because it spends time calculating the occurrences of term t against all term t's it spends a lot of time iterating over each array of tags, in each photo, in each cluster. This means that with only a sample set of 10000 photos split into 100 clusters, it takes several minutes to calculate and output the clusters tags and their respective rankings.

After importing the file into a spreadsheet and sorting the values, it is clear the terms move from generic or high level to very specific or low level. Terms like country/county tags (Ireland, Dublin) are ranked very highly because they appeared very often in this cluster and less so in other clusters. It very quickly drops from broad locations to general terminology, below is a table of the 1<sup>st</sup> 30 tags.

1-10	11-20	21-30
ireland	night	people
architecture	sky	castle
nikon	2008	bw
building	bridge	sign
dublin	street	light
2010	blue	2007
2009	cam:panasonic=tz5	club
church	europe	party
snow	holiday	road
city	music	lights

Given we already know the location of the cluster from its coordinates and the highly rated 'nikon' which as previously discussed is likely a user who's uploaded many photos in this area with that same tag repeatedly, the top 7 tags aren't particularly useful for travel based associations or predictions at first glance. However, most of the other tags are fairly descriptive and show potential.

Lower down the list and towards the bottom are the less common tags, some which have obviously been spelt incorrectly, joined words, coordinates (would only ever be specific for one photo) and also words that were possibly very common across all clusters they ended up with a lower score as a result (tags such as 'from' can be found).

From here it is necessary to retain a certain amount of tags from the list (there are 684 for this cluster) and purge the undesired tags from the cluster so they won't turn up as part of any associations. This will dramatically reduce the number of associations per cluster and it's likely the support (and possibly confidence) threshold will have to be lowered significantly; however any discoveries should be much more meaningful and desirable.

For this test I shall retain the top 100 tags, since I had already imported the chi values for presentation above, I simply removed the top 100 removed the values column, now it is a list of terms to be removed from the cluster – It was necessary to implement another small program that took this list as input and iterated over every term seeing if it was present in each photo and removing it if it was, after all the terms were removed from the cluster was re-output and the same starting parameters were used to find rules:

consequent	antecedent	ante. supp.	support	confidence	lift value	add. eval.
ireland		100.0/112	85.7/96	85.7	1.0	0.0
dublin		100.0/112	93.7/105	93.7	1.0	0.0
ireland	birthday	11.6/13	11.6/13	100.0	1.167	0.0
dublin	birthday	11.6/13	11.6/13	100.0	1.067	0.0
ireland	2007	12.5/14	12.5/14	100.0	1.167	0.0
dublin	2007	12.5/14	12.5/14	100.0	1.067	0.0
ireland	2008	11.6/13	10.7/12	92.3	1.077	0.0
dublin	2008	11.6/13	11.6/13	100.0	1.067	0.0
ireland	2006	12.5/14	12.5/14	100.0	1.167	0.0
dublin	2006	12.5/14	12.5/14	100.0	1.067	0.0
ireland	night	12.5/14	11.6/13	92.9	1.083	0.0
dublin	night	12.5/14	12.5/14	100.0	1.067	0.0
ireland	drinks	17.9/20	17.9/20	100.0	1.167	0.0
dublin	drinks	17.9/20	17.9/20	100.0	1.067	0.0
dublin	gig	14.3/16	14.3/16	100.0	1.067	0.0
dublin	ireland	85.7/96	84.8/95	99.0	1.056	0.0
ireland	dublin	93.7/105	84.8/95	90.5	1.056	0.0
dublin	birthday ireland	11.6/13	11.6/13	100.0	1.067	0.0
ireland	birthday dublin	11.6/13	11.6/13	100.0	1.167	0.0
dublin	2007 ireland	12.5/14	12.5/14	100.0	1.067	0.0
ireland	2007 dublin	12.5/14	12.5/14	100.0	1.167	0.0
dublin	2008 ireland	10.7/12	10.7/12	100.0	1.067	0.0

It's apparent that the rules have changed and some elements are no longer present, such as 'ucd' and 'tkd', I quickly searched for these in the 100 tags that were retained and it appears they were below the threshold so won't appear in any rules for this cluster after feature selection. I lowered the support from 10-100% to 5-10% to again assess the rules:

consequent	antecedent	ante. supp.	support	confidence	lift value	add. eval.
drinks	return	6.2/7	6.2/7	100.0	5.6	0.0
ireland	return	6.2/7	6.2/7	100.0	1.167	0.0
dublin	return	6.2/7	6.2/7	100.0	1.067	0.0

And between 4-5%:

oneills	-6.260758	4.5/5	4.5/5	100.0	22.4	0.0
return	oneills	4.5/5	3.6/4	80.0	12.8	0.0
drinks	oneills	4.5/5	3.6/4	80.0	4.48	0.0

3-4% was not possible because the program ran out of memory – the amount of rules and calculation time grows exponentially as the support gets smaller. Between 5-10% 600 rules were discovered and 4-5% returned 12000+.

Looking at the rules themselves, a majority are still links between the general geographic area and other tags (Dublin and Ireland in this case), although the lower support results do show some more specific items – ‘Oneills’ is a popular bar in Dublin which confirms its association with ‘drinks’, this is by far the most interesting result being a very specific location associated with a decently descriptive tag. Although less support returns less definitive rules, there seems to be a trend across the recent results that show a lower support return associations between more distinct and specific terms.

### 50,000 Photos - 250 Clusters

Ideally I would have used 1000 clusters for this sample size; however my implementation seems to struggle as cluster amount increases. Below is an Argui output of the largest cluster – cluster 230 with 1165 photos:

consequ...	anteced...	ante. su...	support	confiden...	lift value	add. eval.
scotland	uk	11.4/133	10.8/126	94.7	3.049	0.0
edinbur...	uk	11.4/133	10.6/124	93.2	1.679	0.0
edinbur...	scotland	31.1/362	29.6/345	95.3	1.716	0.0
edinbur...	uk scotl...	10.8/126	10.6/123	97.6	1.758	0.0
scotland	uk edin...	10.6/124	10.6/123	99.2	3.192	0.0

Similar to before, the highly supported rules are those with popular generic terms typical of a geographically broad cluster. To test the theory that lower support can reveal more specific, below are an output of rules the top rules with support between 2-3%:

consequent	antecedent	ante. supp.	support	confidence	lift value	add. eval.
unitedkingdom	gbr	2.1/24	2.1/24	100.0	14.747	0.0
geotagged	gbr	2.1/24	2.1/24	100.0	10.13	0.0
edinburgh	gbr	2.1/24	2.0/23	95.8	1.726	0.0
edinburgh	2009	2.2/26	2.0/23	88.5	1.593	0.0
festival	fringe	2.6/30	2.1/25	83.3	31.317	0.0
fringe	festival	2.7/31	2.1/25	80.6	31.317	0.0
edinburgh	fringe	2.6/30	2.6/30	100.0	1.801	0.0
edinburgh	festival	2.7/31	2.6/30	96.8	1.743	0.0
scotland	travel	3.2/37	2.7/31	83.8	2.696	0.0

It's quickly noticeable that the top rules are dominated by appearances of an event 'fringe' and 'festival' – a very popular event in Edinburgh. Perhaps a small optimisation is to implement a maximum threshold to stop incredibly prevalent tags from appearing in associations to allow less supported (but highly confident) rules to be discovered.

Feature selection was attempted with this sample however there are almost 500000 tags that are being iterated over and compared and leaving the program running over night still returned no rankings.

## Further Considerations

During the analysis there were some potential thoughts and ideas I had regarding the project that weren't appropriate to document at that given point in the report. It's obvious from the report and analysis above that extracting a meaningful set of rules can be a tedious task of trial and error, finding suitable thresholds or clusters sizes will change and depend on the sample size so an efficient way to approach rule mining would be to have as many automated optimisation techniques as possible.

Feature selection was used to promote tags that occurred frequently (in many photos) in a given cluster and less frequently in other clusters, this provided a list of ranked tags that was used to remove undesirable tags from consideration. Feature selection itself is the act of discovering desirable features in a dataset and removing the unwanted counterparts. Although this project made use of the Chi-Squared algorithm, I have some more specific types of feature selection that would be beneficial within this context.

**Sentiment analysis** – During my analysis I noticed terms such as 'happy' and 'smile' and although meaningless on their own, coupled with a place or event it might seem like a more promising suggestion for someone rather than if the association was 'bad' or 'unhappy'. Sentiment analysis aims to analyse a given sentence of text and return a value between -1 and 1 to describe how positive or negative the string is. As there are no sentences and only individual words, it would be possible to use a sentiment dictionary (Provalis Research) to rank each tag and give each rule an additional comparative value by summing the values of the individual items. This would of course need testing but in theory it could help further optimise meaningful discoveries by returning more positive rules.

**Timestamp** – like the user ID that was removed at the beginning of the implementation, the timestamp of when the photo was taken was deemed unnecessary and also disregarded. During the analysis it has shown that a date or period of time when the photo was taken (January, Spring, Easter etc.) can be useful for distinguishing when a popular rule containing some event has taken place. The timestamp for each photo could have been used to identify when it was taken and automatically assign the photo with a corresponding tag of choice – it could be a date, month or period, whichever is relevant at the time.

## Conclusion

As an overarching conclusion for this project, I feel it has met its target goal - to implement association rule mining on a Flickr dataset and surmise whether rules can be derived that could help location based services.

However there are many different factors that decide the quality of rules that are discovered and these are what the project conclusion shall document.

## Implementation

The implementation is the least important of the considerations, as there is a fairly clear-cut route for association rule mining, the implementation of the K-Medoids clustering algorithm isn't adjustable and neither is the Chi Squared algorithm. It would of course be possible to change to a different method of clustering; I would suggest that clusters datapoints into geographical regions rather than density would likely be more beneficial for location based services and any rules derived within a cluster could have their clusters origin traced.

Feature selection certainly helped to isolate the most prominent tags within a cluster, however the result is still a judgement, meaning the threshold for cutting off tags by their score is probably going to come from trial and error and if the clusters were large enough, it would change from cluster to cluster. It seems to be a trade-off between a manual intervention to optimise and getting the average from picking a threshold and sticking to it, it's very much an effort versus reward scenario.

In terms of this dataset, manual intervention was only necessary to understand how different levels and thresholds affected the types of discoveries, in a real world scenario it would only be the discoveries themselves that matter, so a manual intervention is only likely if output isn't as expected or desired.

## Analysis

After analysing many of the different rules returned from this project, it is apparent that there are many different types of rules that will benefit different parties in different ways:

Location Based Services, i.e. a holiday recommendation service, an event finder service, tourist information service. All of these can benefit from harvesting rules from photo metadata.

**Holiday Recommendation Service:** Firstly, assuming a holiday is recommended based on something the customer wants, i.e. a service, activity or otherwise, the methodology for association rule mining in this project would be perfectly suitable, that is to say clustering a photoset by density and deriving a rules list for each cluster, then searching through the acquired rules for terms that match the customer's request. A quicker way to do this on a very large dataset would be to skip the clustering and treat the entire initial photoset as one large cluster, derive the rules and then search for the customers term – any rules where the customers term is associated with a location could potentially be used.

**Event Finder Service:** This type of service could easily be implemented as a lightweight application (possibly as a smartphone app) that simply has access to a large database of harvested rules from a massive photoset, the clusters would need to be geographically

identified and similar the holiday recommendation, it would be suitable to simply search for a user's input terms within rules. It would be possible for multiple terms to be searched for considering rules can have a set of items in the consequent – the results could simply return the rules with the most matched terms and because the locations of the clusters are apparent, it could return closer 'events' first as it could search clusters relative to a user's location.

**Tourist Information Service:** This type of service could utilise rules in a slightly different way by classifying popular rules for a cluster into 'events', 'locations' and 'activities' and presenting the top results (with respect to support and confidence) for a given cluster.

**Other Uses:** Another potential use for these rules is within Flickr itself, it could use association rule mining to find strong rules to help predict what tags are likely to appear together (context is more or less irrelevant), this could be used as a service for tag suggestion when users are uploading and tagging their photos, having suggestions for other tags presented (That are found in rules with current tags).

There is a specific paradigm that makes a rule more desirable, and that is one that contains an antecedent that represents a location or activity with a consequent that has multiple items, preferably a location, date/period identification and other related tags, having multiple descriptive tags as a consequent helps understand the semantics of a given rule. The date/period associations are specifically useful for events, as they can be typically annual.

## Reflection

### Summary

Overall I feel the core analysis of the project was a challenging but rewarding experience, I feel my programming knowledge and capability was tested, my intuition and innovation regarding an unknown problem was tested and also my ability to suitably commit time and concentration to a large scale project.

The project itself has proven that structured knowledge can be gained from association rule mining on a large scale photo dataset and this knowledge could potentially be utilised by differing parties. The goal of this project developed from implementing association rule mining on a Flickr dataset into a reflection of methods, techniques, how optimisation at different stages can affect the results and what the most suitable results we could hope to find might be.

### Preparation

The reason this project had such room for expansion/contraction is that each step had deliberate leeway built in to the plan of the project – assisted by the research from the

interim report and insights gained from my supervisors existing knowledge about association rule mining, clustering and feature selection. This made it very clear what implementations would be required to produce results for analysis.

## **Implementation**

All implementation was done in Java and ranged from simple input/output to clustering and feature selection algorithms. This was the most challenging aspect of this project with regard to difficulty; my programming skills can be considered novice at best.

## **Personal accomplishments**

My personal understanding and confidence programming in Java has increased dramatically over the course of this project, the differences can be seen between my 1st program (simple input/output), 2nd program (clustering) and the 3<sup>rd</sup> program (feature selection).

The extent of my programming exposure has been on this University course and nothing that compares to this project with regards to difficulty. Although the idea of object oriented programming and design has been taught in theory, it wasn't until my final program that it seemed logical to read the input into a class of clusters, each with an array of photos (also a class). Then defining functions within these classes to access the data stored inside them upon initial input. It made accessing photo data a much simpler concept when iterating over the dataset.

## **Improvements to the implementation**

As previously stated, my implementations became neater and more efficient as the project progressed, this in turn meant that although my later code worked better, analysing a new sample meant starting at the beginning of the programs, passing each new output to the next program, this became quite disjointed and certainly wasn't an elegant final solution.

Given the chance to re-implement, it would of great benefit to create 1 program with each of the stages in this project implemented as a different function, passing the dataset between them at the first and only runtime. Not only would this save time on outputting & re-inputting data into arrays to duplicate necessary values that the previous implementation had already calculated, it would create a much more satisfying final product.

The eventual product of the implementation was the results that were analysed during this report. Everything that could be potentially analysed had to come from the implementation which is another reason it would have benefitted from an overhaul – changes in the first program would have caused issues with the other programs when trying to interpret the output. They key example of this was the decision to disregard user ID and the Timestamp for each photo during the creation of smaller samples, it later turned out it could have been useful to have these attributes but it would have required changing the initial program and ultimately the rest of them as well.

There was some manual testing of the implemented algorithms – however it was a tedious process to serve a small purpose. The sample reader was tested by manually assessing the photos collected and checking the coordinates to see if they fell within the UK border. The K-medoids was tested by exporting multiple clusters and plotting them on a world map to show their distribution. The Chi Squared algorithm was not checked but the types of tags it returned with a high value compared to those with a low value were within expectations.

I'm sure the implementation would have been more robust if a testing plan had been documented before the start of this final report, this could have helped normalise all the implementations and give peace of mind that everything is working correctly – allowing me to focus solely on analysis.

## Analysis

### My approach to analysis

The analysis for this project went as expected, one obvious omission was the analysis of a dataset larger than 50,000 photos, however the pattern of analysis would have been identical to the other 3 samples.

I feel that my understanding of how various factors (support, confidence, cluster size, sample size) affected the types of discoveries was decent and documented well; furthermore I feel there was a good elaboration on which types of discoveries may be conducive to the project goal.

My approach to changing the variables that affect the nature of the discoveries should have been better structured, by that I mean I should have experimented with a predefined change for all sample sets to get a truer comparison, rather than continually changing the values for each sample until more meaningful results are found – however I do think with any dataset that uses association rule mining to find rules will need to use a certain amount of trial and error.

### Improvements to the analysis

The analysis for this project became somewhat compromised when I realised I had overlooked that fact that Argui can't analyse multiple transaction files at once (clusters in this case). From here it was a decision to implement a major component or make the best of a setback by compensating with a lot of manual interaction. It wasn't really feasible to sacrifice so much time to implement such a large component of the project, additionally the project goal was to assess if and how suitable discovered rules would be for a third party – this turned out to be perfectly attainable from analysing individual clusters. However it would have been additionally beneficial to compare results from all clusters or to aggregate the top rules from each cluster and have a list of the top rules from the entire dataset.

My second gripe regarding my analysis of the various results was that it was such a small and constricted analysis – it would have possibly been more revealing to compare discovered rules in different regions (other than the UK) and possibly compare the tagging habits from different regions.

It would have also been massively beneficial to have organised to use the schools supercomputer to calculate the clusters and Chi Squared values for a much large dataset; it would've meant healthier results and potentially new discoveries.

## Conclusion

Overall I feel satisfied with the conclusion to the project; yes, association rule mining could be utilised to theoretically improve location based services. It would have been a greater accomplishment to identify real world companies or services (other than Flickr) that could have benefited - the conclusion could have been tailored to a more relatable entity.

## Improvements to the conclusion

There are 3 additions I would make to the conclusion if I was to restart the project:

- There were no exact figures given during the conclusion – or at least an investigation to see if changes of variables in a small dataset scaled up and caused the same mutations of rules when changed on a larger dataset and if so, listing some definitive figures.
- There was never a massive dataset tested – when trying to output and cluster a million photos, the program originally suffered a memory leak and ran out of stack space, I couldn't find the problem with the code (since it ran perfectly for other sample sizes) so decided to move forward with the project – considering I was only able to analyse 1 cluster at a time manually, there wasn't much scope for a huge revelation given it was just going to be a cluster with a few more thousand photos inside. Also, given the time that feature selection took to run on 10000 photos, it probably wasn't feasible without use of the school's super computer to run the calculations.
- Lack of different feature selection techniques – Chi Squared was likely the most appropriate given tags are of qualitative nature and this returned a value that allowed tags to be compared – it considered the total distribution of a tag and this helped removed generic tags. It would of however been better for the project conclusion to suggest different feature selection techniques and surmise how each one affects the rules discovered.

A major personal revelation at the end of this project is an understanding for just how significant and beneficial Association Rule Mining can be. There are many successful

companies that make key decisions based on the knowledge gained from ARM – a common example is Tesco’s monitoring customer shopping habits and placing items frequently purchased together near each other in store. However I can imagine there are a whole host of companies with access to transaction like datasets with useful undiscovered knowledge.

Car manufacturers could for example use it in two ways – they could assess customer purchasing habits and make decisions on spending habits similar to Tesco, or they could assess internal carpart faults to identify potential causes or connections between internal mechanical issues. A school or university could treat a student as a transaction and look for correlations between academic results and other attributes. A historical expert could treat historical events as transactions and mine for rules to help predict likely future trends or events.

Of course the above requires a suitable dataset is available and in some cases it won’t be – however the potential ARM has should be enough for companies that could collect data with potential to do so.

Upon rereading my reflection it seems I’ve been rather critical of the project in general, this is obviously because there is enough material to be critical of, however it mainly stems from a personal determination to make note of oversights, ideas or theories as the project advanced and even though a majority could have benefited the results and analysis, it would have probably unfeasible given the time they were detected and the scale of the project already – for the time committed, I feel the project goals were more than satisfied, interesting knowledge was discovered and suitable/practical theories regarding the knowledge were made.

## References

- Abeel, T. (2006). <http://java-ml.sourceforge.net/api/0.1.5/net/sf/javaml/clustering/KMedoids.html>. Retrieved 2013, from <http://java-ml.sourceforge.net>.
- Google. (2008). <https://developers.google.com/kml/documentation/>. Retrieved from [www.google.com](http://www.google.com).
- Mirkes, E. (2011). *K-means and K-medoids applet*. Retrieved from University of Leicester.
- O, L. V., S, S., & B, D. (2012). Georeferencing Flickr resources based on textual meta-data.
- Provalis Research. (n.d.). <http://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/sentiment-dictionaries/>. Retrieved from <http://provalisresearch.com>.
- Simon, A. (2007). <http://processtrends.com/Files/MapExcelData.zip>. Retrieved from <http://processtrends.com/>.