Individual Final Year Project

Initial Plan	–	Acquiring structured knowledge from Flickr
Author	–	Alex Thomson
Supervisor	–	Steve Schockaert
Moderator	–	Alun D Preece
Module Number	–	CM0343
Credits	–	40

Project Description

Flickr is a large photo sharing website where users can upload, label and share photos. These photos are stored in Flickr's database with their respective metadata - unique photo id, owner of the photo, latitude, longitude, tags, time stamp, and the Flickr accuracy level. Flickr have a public web API (Application Programming Interface) which allows users to send requests to Flickr's database to extract collections of data about photos.

Association Rule Mining (ARM) is a widely used technique for analysing and identifying correlations within large datasets. Businesses are always looking to improve their services in a cost effective manner, analysing data they already possess to provide useful information about improving user experience and gaining an advantage over their competition makes ARM a very popular technique. Amazon.com utilises this technique in a simple manner by tracking items purchased by users and identifying whether there are any common similarities between purchases on a large scale, i.e. users that purchase an MP3 player may also quite frequently purchase headphones. Amazon can then utilize this knowledge to suggest to other users when they are buying an MP3 player that they may also like some headphones.

This project is aimed at the analysis of a large dataset gathered from Flickr utilising ARM, the subsequent findings, their potential implications for Flickr and to identify any other parties the resulting information may benefit. In particular this project aims to look at the geographical location of photos alongside their user defined tags and see if there are any strong links between the two and what the resulting information means.

If this analysis were to provide positive results and there were confident rules discovered within the data, we would expect to see consistent similarities between tags and their location, i.e. photos of a certain location may often contain reoccurring tags. This analysis can be expanded to include clustering of the locations in different ways, for example do we get different sets of rules when we cluster photos into counties instead of smaller locations. Perhaps grouping photos like this might contain reoccurring tags that are events, allowing us to plot popular events from county to county.

Hopefully the results from this project will consist of several strong rules extracted from the correlations between user defined tags and the geographical clustering of photos – at various depths of clustering, i.e. buildings, villages, towns, counties etc. Such positive results could be beneficial for many parties:

- Automated or Suggested tagging from Flikr – Providing the results create a taxonomy between tags, Flickr could use the resulting links between words to identify which other words should likely be assigned to a photo
- A tourist recommendation engine could utilise the knowledge about what popular tags occur when the locations are clustered in to counties and then filter the results to identify the popular events at these counties.

Project Aims and Objectives

Objectives

Core

- Research and understand Association Rule Mining comprehensively – ARM is a very complex technique that can be executed on datasets in many different ways, understanding how to use this technique to extract the best possible results is the basis of this project. Researching available literature to grasp how ARM can be utilised with other techniques to analyse place semantics and tag distribution.

- Implement a Clustering Algorithm – A clustering algorithm will be critical to selecting a dataset that will be manageable for whichever tools and methods are used to analyse the data but also able to produce robust results.
- Implement a Feature Selection Algorithm – Feature selection will be necessary to trim unnecessary metadata from the photos, resulting in cleaner and more desirable tags when later analysing for correlations.
- Produce a taxonomy using the tags – analyse the available tags using widely known techniques (Such as TF IDF) to produce a meaningful relationship between tags.
- Produce analysis of results – After understanding the fundamental techniques and implementing the necessary algorithms to obtain a useful dataset, ARM will be used to identify any correlations and the confidence of any rules extracted from them. This will also include theorising what the results could mean for Flickr and any others parties the information may be of benefit to.

  Investigating the scale effects of different clusters will give a further degree by which to measure the usefulness of ARM and give potential to compare the types and confidence of rules extracted from different levels of data (i.e. towns against counties).

Optional

- Testing identified rules – evaluate how well the rules identified, combined with the previously documented taxonomy can accurately recommend tags.
- Evaluate the potential for expanding the dataset – investigate techniques for parallelising parts of the implementation so analysis could be expanded for a much larger dataset, for the purpose of retrieving a greater confidence in rules.

Interim Report

- A background study on Flickr and the services it offers users, specifically photo organisation and what privileges users have when it comes to labelling, tagging, albums, uploading and searching.
- Background on the Flickr API and what data can be gathered from it.
- A detailed section on ARM and the algorithms behind ARM. How it is utilised within other contexts and how it could potentially be used to investigate data gathered from Flickr.
- Investigate previous similar research and how the findings can be used to influence the direction of the project positively.
- An investigation into the available tools for Association Rule Mining and how extensive they are, what format of input are they expecting and whether they suitable for this project.
- Are there similar techniques that can be deployed instead of/along side? How effective are they?
- Investigation into whether a database is suitable for the project, it may be more suitable for example (depending how the analysis is executed) that the input is simply a comma delimited file.
- Research Hypothesis – Theorise about the types of expected results and their implications based on the current knowledge base.

Final Report

Potentially, if required:

- Build database to hold initial data and queries.
- Design and implement a tool to organise data gathered from the Flickr API into easily extractable subsets of data (by way of queries).

Definite Deliverables

- Identify manageable dataset by implementing various clustering algorithms.
- Investigate and implement feature selection techniques on the dataset to retain only desired tags.
- Perform analysis on dataset using identified ARM algorithms.
- A detailed analysis of the results produced, how significant they are and whether they match any initial predictions.
- Validation of results – if there are similar tools available, a comparison of results to determine accuracy.

- A theoretical contemplation of how strong rules could be used to assist businesses, specifically Flickr by way of organisation, searching and suggested tagging.
- A prediction using the knowledge gained from the project as to where and how prospective datasets could be located.

Work Plan

The timescale for the project spans two semesters and has two main deliverables, the interim report (due at the end of semester one) and the final report (due at the end of semester two). It is estimated that the project should take 400 hours to complete (10 hours per credit), considering the interim and final report are worth 25% and 70% respectively, it seems reasonable to spend roughly 100 hours on the interim report and 300 hours on the final report. Compensating for this unevenness in weighting and that the interim report won't be started until week 4 (due to this initial plan), I intend to allocate roughly ¼ of the whole time remaining for this project (26 weeks including Christmas and Easter recess but excluding examination periods) for the interim report and the rest on the final report. This means completing the interim report around week 10 of the first semester and starting the final report over Christmas recess (Which coincides with the submission dates).

The work plan will also reflect time allowances for significant personal commitments I am aware of, the physical writing of the deliverable documents and revision time before examination periods.

For the detailed work plan, please see appendix A.