GENERAL DESCRIPTION OF THE DATASET

Using Wikipedia List of fashion topics (http://en.wikipedia.org/wiki/List_of_fashion_topics) LUH selected 470 fashion topics, out of 1,650 possible topics. The 470 topics were selected because the related categories contained the terms "fashion" or "cloth".

Examples of fashion topics:
A crawler was implemented in Java to retrieve Flickr photos related to the selected fashion topics, using Flickr's API[1] and the Java library flickrj[2].
The crawler gets a set of text queries from a text file and iterates through them to fetch relevant images and metadata from Flickr web services. Each row in the text file corresponds to a fashion topic.
The procedure for crawling is the following:

a. issue a query containing the fashion topic to the flickr search api
b. retrieve the social data, and the actual photos for the first 1000 results and store the information in a database. The actual photo is retrieved with its original size, and if not available with the flickr standard medium size.

For each selected fashion topic we query Flickr for the relevant images, which we retrieve along with all the associated metadata. Thus, for each query we have a list of Flickr photos and their rank according to Flickrs relevance algorithms.
Aside from the actual physical photo, each Flickr photo has a set of associated metadata: information describing the photo, geo-coordinates, tags, favourites, comments, urls, contexts or notes.
The basic metadata information of a photo consists of: Flickr internal id, title, url, owner, description, date taken, date added, date uploaded, date posted and license.
The *license* represents the type of Creative Commons License associated to the photo. The possible licenses attached to a photo on flickr (http://www.flickr.com/services/api/flickr.photos.licenses.getInfo.html) are :

- 4 Attribution License
- 6 Attribution-NoDerivs License
- 3 Attribution-NonCommercial-NoDerivs License
- 2 Attribution-NonCommercial License
- 1 Attribution-NonCommercial-ShareAlike License
- 5 Attribution-ShareAlike License
- 7 No known copyright restrictions

Each photo can have an associated set of geo-coordinates consisting of latitude and longitude

Tags are like a keyword or category label. Tags help flickr users find photos and videos which have something in common. A user can assign up to 75 tags to each photo or video. The tag information retrieved for a photo consists of a list containing tag(id, raw and processed) and the user(id, name) that produced it.

A photo can be declared by one or more users as a favorite. For each photo, when available, we retrieve the list of users(id, name) that declared it as a favorite, and the time when this action occurred.

Users can comment on photos. For the photos that elicited such comments we retrieve the list of users(id, name), the creation date, a link to the comment, and the text of the comment itself.

[1] http://www.flickr.com/services/api/
[2] http://flickrj.sourceforge.net/

Flickr stores the images in different sizes, and provides urls where these can be accessed. For each photo we retrieve the urls of the small, medium, original and large photo.

A photo can appear in one or more contexts. A context is either a group photo pool or a photoset. Users can organize themselves in groups and share photos in a common group pool, for example the group “CeBit 2012” would have a group pool where the members share photos taken at the respective event. Users can organize their own photos in photosets, containing photos that share the common characteristics, for example the “Vacation 2012 South Africa” photo set. For the photos that appear in a certain context, we retrieve the type of context and its id and title.

Users are offered the possibility to create a bounding box on a photo and attach a note to it, enabling them to label or comment only regions of a photo, for example tag persons in a group photo. For the photos that have notes attached we retrieve its id its author(id, name), the coordinates of the bounding box, and its text.

The crawler retrieves images and the associated metadata for a list of fashion topics to be used as queries, specified in a text file, and storing the gathered information in an Oracle database.

Technical details

In the **Metadata** folder you can find the following .csv files, each containing data about a certain type of metadata.

These files were generated as a database export using sqldeveloper.

They can be used directly to be imported into any database.

queries.csv

The columns of the .csv are: SEARCHQUERY, CRAWLTIME, RANK, ID

The *SEARCHQUERY* corresponds to the fashion topic. Each fashion topic has a list of crawled photos identified by their flickr *ID*. The ranking is also stored, represented by *RANK*

info.csv

The columns of the .csv are: ID, TITLE, URL, OWNER, DESCRIPTION, FAVOURITE, ISFAMILY, ISFRIEND, ISPUBLIC, ISPRIMARY, DATETAKEN, DATEUPLOADED, DATEPOSTED, DATEADDED, NTAGS, NCOMMENTS, SECRET, O_URL, ORGINALSECRET, LICENSE

Each photo has some basic associated information.
ID represents the associated flickr internal id, and it is the same as in the *queries* file.
OWNER contains the flickr username of the person that posted this photo
ISFAMILY, *ISFRIEND* reffers to the relation of the user logged in to the *OWNER*(not relevant in this case because it represents the relation of the OWNER with the user we used for the crawling)
ISPUBLIC informs us on the status of the photo
FAVORITE is the number of times the photo was favorited by other users
TITLE, URL, DESCRIPTION, DATETAKEN, DATEUPLOADED, DATEPOSTED, DATEADDED are self explanatory fields
NTAGS, NCOMMENTS are the number of tags and comments respectively
SECRET, O_URL, ORIGINALSECRET are fields necessary for downloading the photo through the API
The *LICENSE* represents the type of Creative Commons License associated to the photo.

geos.csv

The columns in the .csv file are: PHOTO_ID, LATITUDE, LONGITUDE, ACCURACY
This table contains information about the photos that also had geo-coordinates attached to them.
The fields are here self explanatory: *LONGITUDE*, *LATITUDE* represent the coordinates of the photo that has the

flickr id *PHOTO_ID*. The geolocation has a certain *ACCURACY*

locations.csv

The columns in the .csv are: PHOTO_ID, LOCATION
This file contains the location of the photo on the disk.
LOCATION is the full path to the jpg photo inside the folder that keeps the photos.

urls.csv

The columns in the .csv are: PHOTO_ID, SMALL_URL, MEDIUM_URL, ORIGINAL_URL, LARGE_URL
This file holds the locations where the different sizes of photos are stored on the *flickr* servers.
SIZE_URL, where SIZE can be *SMALL, MEDIUM, ORIGINAL, LARGE* are the urls where the photos can be found on the servers of flickr, depending on the size desired. Not all sizes are always available.

contexts.csv

The columns in the .csv are: PHOTO_ID, CONTEXT_TYPE, CONTEXT_ID, CONTEXT_TITLE
A photo can belong to a group pool, or to a set. This file holds these associations.
CONTEXT_TYPE tells us if the context is a group pool or a set
CONTEXT_ID is the identifier of the context
CONTEXT_TITLE the title of the context, to have an idea about what it is

favorites.csv

The columns in the .csv are: PHOTO_ID, PERSON_ID, PERSON_USERNAME, FAVEDATE
Each photo can be designed as favorite by one or more users. This file contains the data about the users that have marked photos as being favorites.
PERSON_ID, PERSON_USERNAME identifiers for the user that assigned this photo as one of his favorites
FAVEDATE the date when this assignment happened

tags.csv

The columns in the .csv are: PHOTO_ID, TAG_ID, AUTHOR_ID, AUTHOR_NAME, TAG_RAW, TAG_TEXT
This file contains the tags associated to a photo. Each photo can have one or more tags, coming from users of flickr.com
TAG_ID the id of the tag
AUTHOR_ID, AUTHOR_NAME identifiers of the user who tagged the photo
TAG_RAW, TAG_TEXT unprocessed and processed text of the tag

notes.csv

The columns in the .csv are: PHOTO_ID, NOTE_ID, AUTHOR_ID, AUTHOR_NAME, X, Y, W, H, NOTE_TEXT
Flickr users have the possibility to mark rectangles on the photo and add some text to these rectangles. For example if you have a photo with a complete outfit, each individual item may be marked with a rectangle bounding box, and have associated a text describing it. This file contains the notes associated to our photos.
NOTE_ID the identifier of the note
*AUTHOR_ID, AUTHOR_*NAME identifiers of the user who posted this note
X, Y, W, H the coordinates and dimensions of the rectangle that contains the note
NOTE_TEXT the text of the note

comments.csv

The columns in the .csv are: PHOTO_ID, COMMENT_ID, AUTHOR_ID, AUTHOR_NAME, DATECREATED, PERMALINK COMMENT_TEXT
Flickr users can also comment on all the photos. This file contains the comments associated to the photos in our crawl.
COMMENT_ID the identifier of the comment
*AUTHOR_ID, AUTHOR_*NAME identifiers of the user who posted this note

DATE_CREATED the date when the comment was made
PERMALINK a link where the comment can always be accessed
COMMENT_TEXT the text of the comment

Statistics

Pairs fashion item, photo: 4835
Number of distinct fashion items : 154
Number of distinct photos: 4810
Max/avg/min nr of photos per fashion item: 531/ 31.40 / 9

Number of photos with geo annotations : 1547

Total number of comments: 9501
Max/avg/min nr of comments per photo: 464 / 7.16/ 1

Total number of tags, photo pairs : 83128
Total number of distinct tags: 11691
Max/avg/min nr of tags per photo: 75/ 17.87/ 1

Total number of notes, photo pairs: 497
Max/avg/min nr of notes per photo: 41/ 3.43/ 1

Total number of favorites: 7265
Max/avg/min nr of favorites per photo: 50/ 4.74/ 1

Total number of contexts: 20185
Max/avg/min nr of contexts per photo: 130/ 4.89/ 1

Top 10 used tags:

No. Apparitions	TagName
633	art
592	red
591	nikon
581	boat
578	water
573	green
569	blue
567	photo
566	trees
566	photography

Dates of photos statistics:
First date of uploading : 28-MAR-05
Last date of uploading: 09-JUL-12

Annotations

In the **annotation** folder you can find the annotation results and analysis from Amazon Mechanical Turk (AMT) for both non-expert and trusted annotators.

There are 4 files in this folder:

MTurk_NonExperts_Results.csv: this file contains the results from AMT for non-expert annotators. This file contains the input and output for each assignment in csv format. Each row corresponds to an assignment which is done by a turker. The field with name "Answer.related(i)" is the answer of turker to question "**Fashion Related**" for the ith image in question (i = 1..4) and the field with name "Answer.specialty(i)" is turkers answer to question "**Specialty clothing item**" for the ith image. For privacy purposes the WorkerId column is replaced with a WorkerIndex column.

MTurk_Trusted_Results.csv: This file has exactly same format as Turk_NonExperts_Results but the results are coming from trusted users (mostly paper authors).

Annotation_PerImage_NonExperts.csv: This file is the answer of turkers to HITs which is calculated based on each image. Below description of each field is provided:

PHOTO_ID: this field is the photo id of the image in this record.

PictureURL: the url of image in flicker.

Category: the specialty category which is the given tag of user to image.

T1_Q1: answer of first turker to question1 (fashion related question)

T2_Q1: answer of second turker to question1 (fashion related question)

T3_Q1: answer of third turker to question1 (fashion related question)

T1_Q2: answer of first turker to question 2 (specialty clothing question)

T1_Q2: answer of second turker to question 2 (specialty clothing question)

T1_Q2: first third turker to question 2 (specialty clothing question)

T1_Familiarity: answer of first turker to familiarity question (how familiar are you with specialty clothing item?) the value is from 1 (unfamiliar) to 7 (familiar)

T2_Familiarity: answer of second turker to familiarity question

T3_Familiarity: answer of third turker to familiarity question

Annotation_PerImage_Trusted.csv: This file has exactly same structure as previous file but the data is derived from trusted annotators.

Majority_Voting.csv: This file contains the majority vote of all annotators (non-experts and trusted annotators) to both questions. Here is the description of each fields:

Url: the url of image

Q(i)ExpertVote: the majority vote of expert (trusted) annotators to question i (i = 1,2). The value can be Yes, No, NotSure and NoAgreement.

Q(i)NonExpertVote: same as previous field but for non-expert users

Q(i)CountYes: number of Yes values for previous two fields. Value would be from 0 to 2.

Q(i)CountNo: same as previous field but for No value

Q(i)CountNotsure: same as previous field but for Notsure value

Q(i)CountNoAgreement: same as previous field but for Notsure value.

Images

In folder **Images** you can find the actual images. There are 154 subfolder in this folder each corresponding to a category. The folder names are same as category names and each folder there are images for that category. There are in total 4835 images.

Related Publication

B. Loni, M. Menendez, M. Georgescu, L. Galli, C. Massari, I. S. Altingovde, D. Martinenghi, M. Melenhorst, R. Vliegendhart and M. Larson. Fashion-focused Creative Commons Social Dataset. *Multimedia Systems Conference*. Oslo, Norway. March 2013.