Multimedia and User Credibility Knowledge Extraction

Deliverable 1.1
Report on Existing Data Collections

Document Information

Delivery date:	01/03/2013
Lead partner:	TUW
Author(s):	Mihai Lupu, Adrian Popescu, Adrian Iftene, Pinar Sahin, Allan Hanbury
Participant(s):	CEA, Bilkent, UAIC
Workpackage:	1
Workpackage title:	Data collection
Workpackage leader:	CEA
Dissemination Level:	PU – Public

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
0.1	30/1/2013	Draft	TUW	First draft
0.9	20/2/2013	Pre-final	TUW	Added in content from UAIC
1	28/2/2013	Final	TUW	Final comments from partners

Abstract

This documents identifies the existing data collections for image search and associated meta-data. It also looks beyond the set of existing collections at tools or resources that will allow us to collect the data necessary for the project.

Table of Contents

1. Introduction

One of the project goals is to prove the feasibility of the models and methods over large-scale multimodal data. To speed up the collection so as to have the necessary data available as quick as possible, a distributed data collection framework will be implemented.

The data collected during MUCKE are highly dynamic and complex, two characteristics which require for the extraction framework to be implemented in a flexible manner so as to cope with new data whenever needed. Information about one million Flickr users, with around 100 million associated images and metadata, and about 10 million Wikipedia articles in 4 languages, with associated multimedia elements from Wikimedia. The concept similarity resources will include around 100,000 multimedia concepts, roughly five times more than ImageNet, a widely used resource in computer vision.

The implementation of the project requires data collection from different Web sources, including user-provided information. All collected user data will come from publicly available sources and will be collected with strict adherence to the data collection policies of the original sources. More generally, all data collection and processing will be performed with strict respect for privacy, copyright or other licensing information.

2. Existing data collections

While the data collection is on-going, the project uses existing data collections, which we list here, together with their summaries and advantages/disadvantages for the purposes of the project.

2.1. ImageNet (www.image-net.org)

ImageNet is a collection of over 14 million annotated images against almost 22 thousands synsets from WordNet (nouns only). The collection of images is made available by either a list of URLs which can then be individually downloaded, or as a collection of original images, after having signed a license agreement. Additionally, the ImageNet project makes available via an API

- SIFT[I] features via an API,
- object bounding boxes
- object attributes

The object bounding boxes and attributes are created and verified via crowdsourcing, but are not present for all synsets available in the collection.

2.2. ClueWeb09/12 (lemurproject.org/[clueweb09/|clueweb12.php])

The ClueWeb09 collection was created by Jamie Callan’s group at CMU in early 2009 and is the test collection now used in the largest tasks of TREC. It consists of 1 billion web pages (25TB uncompressed) but stores only the HTML part of the web pages it has crawled. Nonetheless, images and potentially even CSS scripts can be fetched subsequently through the URLs by which they are referenced in the HTML body.

The collection is distributed as a series of warc[II] files, each warc file containing tens of thousands of pages, for an average size of 1GB of uncompressed data.
The 2012 version of this collection has also been recently released[III]. This collection includes images, but truncates every file larger than 10MB. It still does not contain multimedia files (Flash, video, sounds).

2.3. Wikipedia (dumps.wikimedia.org/)

Full copies of Wikipedia are available as part of both ClueWeb09 or ClueWeb12. Wikitravel is equally part of ClueWeb12. Additionally, Wikipedia can be downloaded directly in a variety of formats from the Wikipedia site itself. However, also in this case, images are not distributed in bulk directly by Wikipedia. Several mirrors provide bulks of images and uploaded files[1], or they can be downloaded separately based on the URLs present in the textual dumps. These textual dumps are in SQL/XML format and are direct dumps of the mediaWiki software. The HTML versions are not regularly maintained (last dump on official website: 2008).
The complete set is about 6TB compressed. A bulk download tool of the latest dumps is available on GitHub.[2]

2.4. Tweets2011 (trec.nist.gov/data/tweets/)

Tweets2011 is the corpus used for IR evaluation in the TREC Microblog track. It contains a list of tweet identifiers (two weeks of tweets), which the user needs to download directly from Twitter using an also provided tool.[3]

2.5. MIRFLICKR (press.liacs.nl/mirflickr/)

1 million Flickr images under the Creative Commons License are available with the MIRFLICKR collection, now used in the ImageClef[4] photo annotation task. The collection is crawled using the Flick API and images are selected by their interestigness factor. In addition to images, the collection contains the user-supplied Flickr tags as well as EXIF metadata. Additionally, for 25 thousands of them, manual annotations are also available.

2.6. CommonCrawl (commoncrawl.org)

CommonCrawl is the largest web collection currently available, 6 billion pages, including all payload information, stored in arc files on the Amazon Cloud. The collection is not meant to be downloaded in bulk, but rather worked on in the cloud.

[1] http://meta.wikimedia.org/wiki/Mirroring_Wikimedia_project_XML_dumps#Media
[2] http://github.com/babilen/wp-download/
[3] https://github.com/lintool/twitter-tools
[4] http://imageclef.org/2012

3. Indirect sources

In addition to the collections described above, an indirect way of reaching the data is through the variety of online image search services. This part of the deliverable gives the reader an overview of such tools.

3.1. Google Image Service

Through Google Image Service the user can manipulate image data using a dedicated Java API. Images can be resized, rotated, flipped and cropped. Multiple images can be composed into a single one. Image data can be converted between several formats. Additionally, photographs can be enhances using a predefined algorithm. The API can also provide information about an image, such as its format, width, height, and a histogram of colour values.

```
import com.google.appengine.api.images.Image;
import com.google.appengine.api.images.ImagesService;
import com.google.appengine.api.images.ImagesServiceFactory;
import com.google.appengine.api.images.Transform;

        byte[] oldImageData;
        ImagesService imagesService = ImagesServiceFactory.getImagesService();
        Image oldImage = ImagesServiceFactory.makeImage(oldImageData);
        Transform resize = ImagesServiceFactory.makeResize(200, 300);
        Image newImage = imagesService.applyTransform(resize, oldImage);
        byte[] newImageData = newImage.getImageData();
```

Code snippet 1: Sample of code using the Google Images Java API to resize an image

3.2. Bing Search

Using the Bing Search API Version 2**,** developers can create applications that retrieve information from the Internet, allowing their users to access Bing's web, image, news, and video search results in formats like JSON, XML or OData, and improve and enhance search requests and results. By signing up, one has 5000 transactions/month for free. For greater amounts of transactions, a certain fee must be paid.

3.3. Picsearch

One interesting aspect of Picsearch is that it indexes images from the web using a web-crawler (Psbot). Using a load spreading technique, it reduces the load on the web-servers of the indexed domains. Unfortunately, it does not provide an API for developers to integrate its functionalities into other applications.

3.4. Yahoo! Search

Yahoo! BOSS Search API is an open search and data services platform based on REST, providing access to web, image, news, spelling and blog Yahoo! Search results with a simple pricing scheme based on usage. It uses OAuth as a simple and secure method for validation and access. This is an open authorization model based on existing standards, ensuring that secure credentials can be provisioned and verified by different software platforms.

3.5. Flickr

Flickr is an image and video hosting website, web services suite, and online community. It boasts more than 6 billion images being hosted on its servers. Flickr provides a filtering system that enables its members to mention the types of the photographs they upload, and also it lets users search for pictures in the same manner. It comes with a complex API, which can be accessed through REST as well as SOAP, with a vast documentation and API Kits for every modern programming language.

```
//initialize SearchParameter object, this object stores the search keyword
SearchParameters searchParams=new SearchParameters();
searchParams.setSort(SearchParameters.INTERESTINGNESS_DESC);

//Create tag keyword array
String[] tags=new String[]{"Dog","Beagle"};
searchParams.setTags(tags);

//Initialize PhotosInterface object
PhotosInterface photosInterface=flickr.getPhotosInterface();
//Execute search with entered tags
PhotoList photoList=photosInterface.search(searchParams,20,1);
```

Code snippet 2: Searching for images by keywords using Java and the Flickr API

3.6. Picasa

Picasa is an image organizer and viewer for managing and editing digital photographs. It is also an integrated photo-sharing platform. For organizing purposes, Picasa comes with file importing and tracking features, as well as tags, facial recognition, and collections for further sorting. Searches can be made by filenames, captions, tags, folder names, and other metadata.

```
URL feedUrl = new URL("https://picasaweb.google.com/data/feed/api/user/Irnuk");
Query myQuery = new Query(feedUrl);
myQuery.setStringCustomParameter("kind", "photo");
myQuery.setStringCustomParameter("tag", "puppy");
AlbumFeed searchResultsFeed = myService.query(myQuery, AlbumFeed.class);
for (PhotoEntry photo : searchResultsFeed.getPhotoEntries()) {
    System.out.println(photo.getTitle().getPlainText());
}
```

Code snippet 3: Code sample that searches images by tags for the user “Irnuk”

3.7. Photobucket

Free alternative to large image hosting, **Photobucket** is also a video hosting, slideshow creation and photo sharing website. It is a good source of great images, but most of them do not have copyright notices, therefore should be reused with care.

Photobucket is mostly used for personal albums, remotely storing avatars displayed on forums, and video hosting. Images store here are often used for eBay, Myspace (from 2007–2009, a corporate cousin), Facebook, LiveJournal, Open Diary - generally blogs and message boards. Users can keep their albums private, allow password-protected guest access, or open them to the public.

Photobucket also provides an API for third party applications, which can be used for uploading images and videos, getting all recent media (videos and images) for a specific user/all users/group albums, searching media matching specific terms in one user's account/all user accounts/ group albums, getting all details associated with one piece of media, such as link URLs/thumbnail URL or updating titles, descriptions, and tags. It uses a unique API key for signing and authenticating by means of OAuth, and REST for requests and responses.

3.8. DeviantArt

DeviantArt is an online community showcasing user-created artwork. It contains over 100 million original works of art made by artists from over 190 countries. The photos are organized in various categories and can be downloaded for free or bought as prints. However, it only provides an API that only allows users to integrate their accounts with their other applications and websites.

3.9. Shutterstock

Shutterstock serves the purpose of selling stock photos, stock vector graphics made in programs like Adobe Illustrator or Macromedia Freehand, and stock raster illustrations created in 3D graphics programs or bitmap editors like Adobe Photoshop. It maintains a library of royalty-freestock photos, vectors, and illustrations available by subscription. Visitors can browse the library for free, and can license and download images through a variety of subscription packages. Thousands of images are added to this library each day as photographers and illustrators from around the world submit their work. Images to be added to the collection are chosen based on a selection process that takes into account their quality, focus, aesthetics, artistry, and originality.

The Shutterstock REST API allows outside developers to create their own tools that make use of the functionality of the Shutterstock website. They can read and update information about the authenticated customer, but more importantly, they can search the stock for images matching specific criteria (category, submitter, language, orientation, author, keywords).

3.10. Open Clip Art Library

The Open Clip Art Library (OCAL) is a large collaborative community that creates, shares and remixes clipart. Its resources may be used in any project for free without any restrictions. As of October 2007, OCAL incorporated over 10,000 images from over 500 artists, the entire collection being available for free download. All images are donated to the public domain by their contributors and are stored in Scalable Vector Graphics (*.svg*) format, often with thumbnails in Portable Network Graphics (*.png*) format.

There is an API available for communicating with OCAL. Requests can be made using REST and it provides information such query engine method, latest comments, top contributors, etc.

3.11. Getty Images

Through Getty Images users can directly access millions of images and hundreds of thousands of videos, associated with an extensive and rich set of metadata and keywords. Files can be downloaded on all the sizes available, from very small ones to high-resolution.

Their API enables integration of the expansive content of the website, making it possible to use the powerful search and rich metadata to create new products and services. For all images matching a particular query, the search is partitioned into a single query and an optional set of filter. This supports scenarios when the client submits a query and the server not only responds with an initial set of results, but also suggests possible modifications to that query. Thus, the users can interactively modify their search following the suggestions that intend to improve the quality of matches.

3.12. Panoramio

Panoramio is a website that makes it possible for digital photographers to geolocate, store and organize their photographs and to also view them in Google Earth and Google Maps. It is a community-powered site that focuses on exploring places (e.g. cities or natural wonders) through photography. The main difference from other image sharing sites is that all the photographs are meant to illustrate places – Panoramio is all about seeing the world.

With the help of Panoramio API, geolocated photos can enrich maps or illustrate information whenever location plays an important role (e.g. real estate sites, hotels and vacation sites, routes and trails). This API uses REST and its responses are formatted in JSON

4. Conclusion

We have shown in Section 2 and Section 3 the current state in terms of existing image data collections, as well as potential intermediaries in generating the new collection. It is clear that there exists no collection at the scale we have planned, while there are several useful sources of data.

[I] David G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 2004.

[II] Web Archive format, WARC ISO 28500 final draft (as of June 18th, 2008), version 018. · http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

[III] http://boston.lti.cs.cmu.edu/clueweb12/