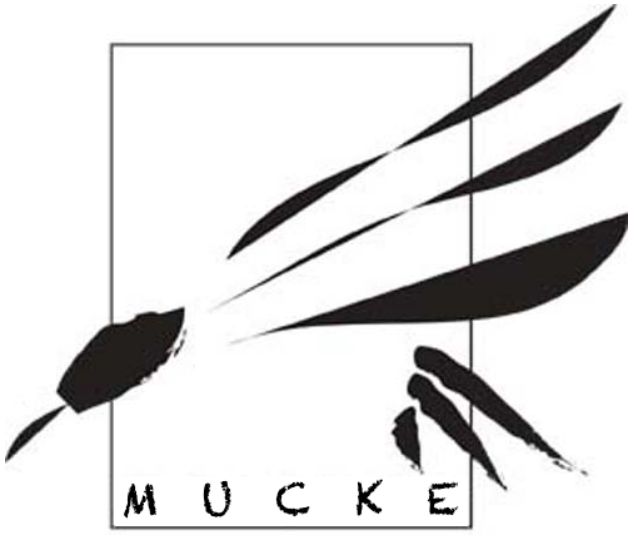

Multimedia and User Credibility Knowledge Extraction

Deliverable 6.3
Report on Resource Sharing Framework

chist-era

Document Information

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

History of Versions

Version	Date	Status	Author (Partner)	Description/Approval Level
0.1	25/06/2013	Draft	UAIC	First draft

Abstract

This document summarizes the existing resources from each partner and describes the prototype built at this moment in the project. The current application allows searching in the local collection of images downloaded from Flickr, or remote, using Flickr public API, in case no relevant results are found in our resources. The main advantage of the prototype architecture comes from the it's scalability and loose coupling, allowing us to add at any moment additional components for text or image processing, or components that deal with user credibility.

The description of WP6 and Task 6.3 taken from project proposal are presented below:

WP6 – Evaluation and Resource Sharing

The objective of this WP is to evaluate the algorithms developed and to propose services dedicated to resource sharing. Algorithms will be tested within existing relevant evaluation campaigns and/or against representative datasets and a new evaluation task will be proposed for user credibility estimation. Resource sharing will be implemented in order to give external interested parties the occasion to exploit a part of the project results.

Task T6.3 - Resource Sharing

Most of the resources created during the project are of high interest for the different scientific communities and they will be in part shared. Visual models will be created for around 100,000 concepts and they are of interest for the computer vision community.

In order to respect copyright, we will adopt a distribution strategy similar to that of ImageNet, in which only links to images are shared in order to avoid copyright issues. These lists of links will be proposed for direct download.

Multimedia concept descriptions will be shared via a Web service which will provide access to the mappings on entry texts to related textual or multimedia concepts from our similarity framework.

The task will be coordinated by UAIC, who will centralize the relevant resources created by all participants and will implement the services necessary for their sharing.
It will take place in two stages: first an internal sharing mechanism (M7-M9) and subsequently an external sharing mechanism (M25-M27).

Table of Contents

1. Introduction

Section 2 of this report presents available resources from each partner and Section 3 describes the skeleton of the application specifically designed to provide access to entry texts and multimedia concepts mappings. Thus, we have performed the following steps: extraction of visual features, indexing, metadata processing and textual processing.

2. Resources

In this section we present all available machines from every partner which can be used in this project (Section 2.1). On these machines, resources and services for text and image processing are available (Section 2.2).

2.1. Machines

In the next table we summarize the computers, clusters, servers and storage resources available at all involved partners:

Partner	Address	Description
TUW	ldc.ir-facility.org	Large Data Collider - old, big server: 80 Intel Itanium Cores, 300GB RAM, about 10TB available in different partitions. Rather unreliable machine (it's from 2007) but accessible from outside our network via SSH. Additionally, we can whitelist external IPs to connect on specific other ports.
TUW	stutomcat.ifs.tuwien.ac.at	Very small virtual machine running Tomcat and accessible from outside on ports 80 and 8080. Used by us to provide access to some utilities via web interfaces. Because it is inside our network it has access to the LDC on all ports, so we have data on the LDC shown to the world via the stutomcat machine. It is here that the download task distribution service runs.
TUW	gonzo.ifs.tuwien.ac.at	Office server, used, among others, to download data. It is only available from inside our office network (i.e. not even from the TUW network).
CEA	Internal address	Two old clusters (2006 and 2009) with: around 100 and 60 Intel cores respectively: 4GB and 16 GB RAM/node. Both are unreliable but still usable to some extent.
CEA - TBD	Internal address	The dedicated MUCKE cluster is due to arrive in September
UAIC	http://metashare.infoiasi.ro/	Server with information about UAIC NLP services. Configuration of this server is Dual Intel XEON X5650, 24 GB Ram, 2 x SAS, 146 GB, 5 x SATA3, 2 TB.
UAIC	http://info-c-12.info.uaic.ro/	Server with information about UAIC NLP services. Configuration of this server is Dual Intel XEON X5650, 24 GB Ram, 2 x SAS, 146 GB, 5 x SATA3, 2 TB.
BU	Internal address	2x128 GB SSD Harddrive 3 TB 7200rpm Internal Harddrive 12 TB 7200rpm External Harddrive
BU - TBD	Internal address	We have 2 workstations with 2cpu (20MB cache for each), 16cores, 32threads, 128 gb RAM and 6 GB Graphic processor. These workstations are not accessible from outside of Bilkent network.

2.2. Libraries, software

In the next table we summarize the resources and tools available at all involved partners:

Partner/ Resource/ Tool	Resource/Tool Name	Description
TUW – Tool – In use	Task Distribution service	For each image list, an application can query the TDS and receives back a task id, and a pair of numbers indicating the start and end position in the list. After the images are downloaded, the application must send a notification to the TDS to indicate this status.
TUW	Astera	Astera is a research project on multimodal retrieval, whose lessons learned will be shared with MUCKE.
CEA – Resource - TBD	Explicit Semantic Analysis (ESA)	Models (inverted indexes) for the four languages modelled (English, French, German and Romanian). Both "classical" ESA and improved ones obtained in MUCKE will be proposed.
CEA – Resource - TBD	Ranked image list	For each of the modelled concepts, a ranked image list is computed, where the automatically obtained rank expresses the probability for that image to be relevant
CEA – Resource - TBD	Textual models for each concept	These models will be obtained through a mapping of the concepts onto a large conceptual space defined by ESA
CEA – Resource - TBD	Visual models for each concept	These models will be obtained by compacting visual descriptions of each concept using techniques inspired from text processing.
CEA – Resource - TBD	Similarity matrices in the textual and the visual domains	These matrices will be obtained by computing concept similarities in the textual and, respectively, visual domains
CEA – Resource - TBD	Similarity matrix in the multimedia domain	Obtained through a fusion of textual and visual similarities.
CEA – Tool – In use	MM - multimedia indexing engine	Tool that incorporates different NLP and image processing components. For text processing, there are morphological and syntactical analysis in different languages, including French, English, and German. For image processing, large arrays of local and global features are implemented locally and the tool is interfaced with OpenCV.
CEA – Tool – In use	ESA (Explicit Semantic Analysis) extractor	Tool that computes "classical" ESA vectors in different languages, among which the four that we promised to handle in MUCKE. Integration of disambiguation and anaphora resolution is now needed.
CEA – Tool – In use	Flickr Download Scripts	The Flickr metadata and image download scripts that are integrated in the distribution system by TUW.
CEA – Resource – In use	Flickr dataset	A dataset of around 3 million Flickr images that can be used to make initial image processing tests.
CEA – Resource – In use	ImageCLEF Wikipedia Retrieval dasets	ImageCLEF Wikipedia Retrieval 2010 and 2011 datasets - http://imageclef.org/2010/wiki, http://imageclef.org/2011/Wikipedia
CEA – Resource – In use	ImageCLEF MIR Flickr datasets	ImageCLEF MIR Flickr 2011 and 2012 datasets - http://imageclef.org/2011/Photo, http://imageclef.org/2012/Photo
CEA – Resource – In use	Credibility Retrieval Dataset	Based on ImageCLEF Wikipedia retrieval datasets, we have built a ground truth for around 90 topics that are well enough represented in Flickr. This ground truth is adapted for evaluating data credibility (quality) since the uploaders of the photos are known and, for each topic, there are 20 images from 15 different users (300 images/topic). All images were manually judged by 3 annotators

Partner/ Resource/ Tool	Resource/Tool Name	Description
CEA – Resource – In use	Disambiguation ground truth	We selected ImageNet images for over 100 concepts that are ambiguous (i.e. that have at least two different associated concepts in ImageNet). This dataset can be used to perform textual, visual or multimodal tag disambiguation.
UAIC – Tool – In use	PoS tagger for Romanian	A hybrid part of speech tagger which successfully combines a statistic model with a rule based system
UAIC – Tool – In use	Graphical Grammar Studio (GGS)	Graphical Grammar Studio is a tool for applying grammars which behave as words acceptors/consumers and annotators. GGS grammars can be used to find and annotate sequences of words which respect certain conditions, in a given input.
UAIC – Tool – In use	NP chunker for Romanian	GGS can create complex grammars which can even work as standalone NLP tools. This tool uses a complex grammar which recursively detects and annotates noun phrase chunks for Romanian text.
UAIC – Tool – In use	Dependency parser for Romanian	This tool determines dependency trees of input Romanian text. The tool is based on the Malt Parser library and it uses Nivre's algorithm.
UAIC – Tool – In use	Clause splitter for Romanian	The tool adds delimitations around the clauses present in the input text.
UAIC – Tool – In use	Discourse parser for Romanian	The tool builds a binary RST-like tree structure of an input text. Nodes in the tree are discourse spans, leaves are discourse units (clauses or simple sentences). Nodes of the tree are labelled as nuclear or satellite. Under any node there are either two nuclei or one nucleus and a one satellite.
UAIC – Tool – In use	Text categorizer for English (CategoriZer)	CategoriZer provides tools for automatic extraction of language indicators that help researches to monitor language use. Here, language indicators refer to: frequent words, metrics, key words, etc.
BU – Resource – In use	ImageNet feature extractor scripts	The ImageNet collection was downloaded. From the Tiny Image Collection features like SIFT and colour features were extracted.
BU – Resource – In use	Yahoo! Large Scale Flickr Tag Image Classification Challenge Dataset	2 million Flickr images downloaded for this dataset.
BU – Resource – In use	Yahoo! Large Scale Flickr Tag Image Classification Challenge Dataset features	Local and Global feature extractor scripts are prepared, including RGB, Lab and HSV Color Histograms, GIST and SIFT features.
BU – External resource	VlFeat	VlFeat Matlab library for feature extractions etc.
BU – External resource	OpenCV	OpenCV Open Source Computer Vision library

3. Internal Sharing Mechanism

The application skeleton was built based on the Lucene[1] application (for the document retrieval part) in combination with LIRe[2] (for the image

[1] Lucene: http://lucene.apache.org/

retrieval part). LIRe is an efficient and light weight open source library built on top of Lucene, which provides a simple way for performing content based image retrieval. It creates a Lucene index of images and offers the necessary mechanism for searching this index and also for browsing and filtering the results. Being based on a light weight embedded text search engine, it is easy to integrate in applications without having to rely on a database server. Furthermore, LIRe scales well up to millions of images with hash based approximate indexing.
LIRe is built on top of the open source text search engine Lucene. As in text retrieval, images have to be indexed in order to be retrieved later on. Documents consisting of fields, having a name and a value, are organized in the form of an index that is typically stored in the file system.
The system was designed with a modular architecture, which will allow us to dynamically integrate new techniques and new algorithms to achieve suitable matches in the future.

3.1. Extraction of visual features and indexing

Using LIRe, we are able to extract, index and search by the following features of raster images:

- **Colour histograms** in RGB (Red-Green-Blue) and HSV (Hue-Saturation-Value) colour space. Colour histograms are a representation of the distribution of colours in an image;
- MPEG-7 descriptors **scalable colour**, **colour layout** and **edge histogram**. MPEG-7 includes standardized tools (descriptors, description schemes, and language) that enable structural, detailed descriptions of audio–visual information;
- **The Tamura texture features coarseness, contrast and directionality**. Six basic textural features were approximated in computational form - *coarseness, contrast, directionality, line-likeness, regularity, and roughness*. The first three of these features are available in LIRe;
- **Colour and edge directivity descriptor (CEDD)**. This feature incorporates colour and texture information in a histogram and is limited to 54 bytes per image;
- **Fuzzy color and texture histogram (FCTH)**. This feature also combines, in one histogram, colour and texture information. It is the result of combining three fuzzy systems and is limited to 72 bytes per image;
- **Joint Composite Descriptor (JCD)**. One of the Compact Composite Descriptors available for visual description, JCD was designed for natural colour images and results from the combination of two compact composite descriptor, CEDD and FCTH;
- **Auto colour correlation feature**. This feature distils the spatial correlation of colours, and is both effective and inexpensive for content-based image retrieval.

In order to create an index and perform searches on it, the following steps are to be followed:

1. For each indexed image in the collection, a *document* is created. This document may contain both textual fields and visual features fields

[2] LIRe: http://www.lire-project.net/

(from those mentioned above). Later, this document is added to the index;
2. To perform searches on the index, a query document first has to be created. This document must contain the fields necessary for the search (textual or visual) – in other words, the search criteria.

The result of the search using the query document is a list of documents with their attached scores in descending order (1 is the best score, while 0 is the worst). These scores illustrate the fitness level between a document and the query document, based on the search criteria. The search process does not have to stop here - the set of results can be further filtered using other criteria.

3.2. Metadata processing and textual processing

Metadata based image retrieval has been widely used over the years due to its simplicity and its low computational cost. Images are manually or automatically annotated with keywords that are stored in databases in order to allow future access to the image.

The Flickr corpus provides a set of associated metadata for each image, which contains information regarding the resource. The most important data provided by the Flickr repository are the fields referring to owner, title, important dates (upload date or create date), localization (GPS coordinates), but also free content such as user tags. These fields retain user input (tags, title) or automatically identified data (GPS location, dates) and represent a key part in an image retrieval task. All this metadata is textually processed. In the first phase we do anaphora resolution and document retrieval using Lucene. In the second phase we intend to use semantic processing, named entity recognition, etc.

However, text annotations often carry little information about images visual features and they are usually associated with subjectivity, ambiguity and imprecision caused by specifying the context and semantic context of images. This leads to the necessity of integrating both content and metadata descriptions for an efficient image data management.

3.3. Prototype Skeleton

Below we will describe the main steps performed in our internal sharing mechanism prototype:

1) The customer enters the query: keyword query mode (content-based or text-based);
2) The server receives keywords and checks in a NoSQL database if there is any table with every keyword:
   2.1 if yes, then it interrogates the NoSQL database and skips to step 8;
   2.2 if not, it adds the keyword in the SQL database tables *Keyword* and *UserKeyword;*
5) The server performs a call to the Flickr API and receives a collection of metadata files;
6) The server applies a clustering mechanism dependant on the query results from step 1;
7) The results are saved in the NoSQL database as follows:
   7.1 A hash keyword called *HashValue* is made*;*

7.2 A table with *cb[HashValue]* or *tb[HashValue]* name, depending on the query (cb = content based, tb = text based), is created

7.3 Saves in table clusters with metadata. PartitionKey represents cluster ID and RowKey represents image ID

8) Return to the client cluster to display the results

Our collection is based on a Flickr[3] collection of images. In the prototype, a *search page* allows inserting keywords for search process (see Figure 1). First, the keywords are searched in image titles or image description fields. Second, based on the content of the images we create a cluster of similar images (see Figure 2).

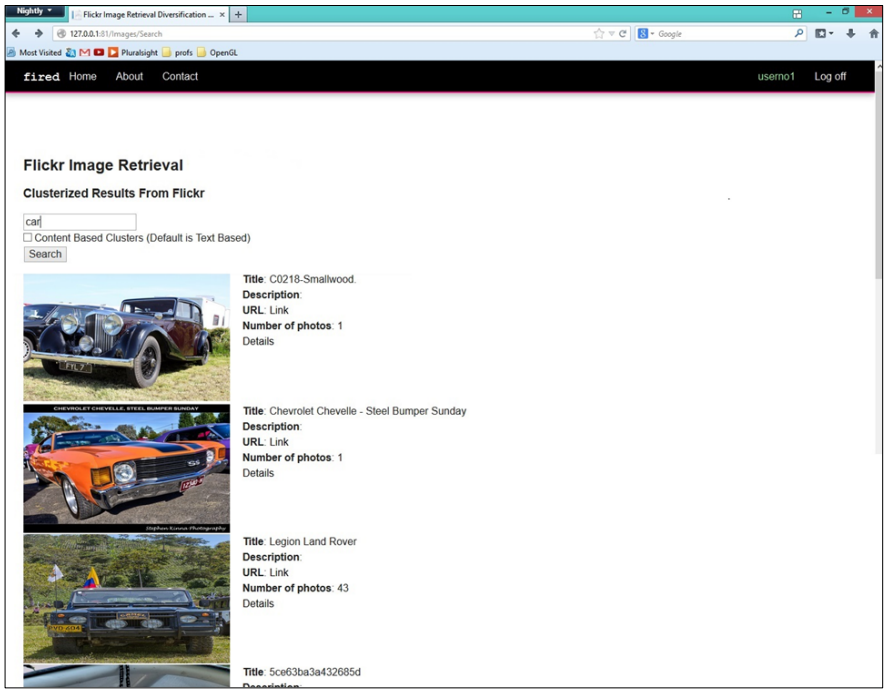

Figure 1. Flicker Image Retrieval Prototype

Clustering based on image content page – In Figure 2 we can see one cluster from the above search.

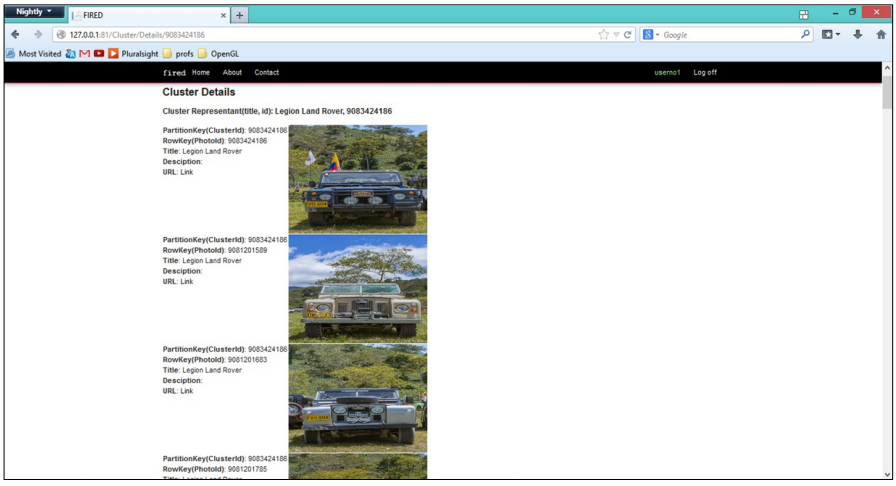

Figure 2. Cluster based on Image prototype

[3] Flickr: http://www.flickr.com/

The main advantage of this architecture is its modularity, enabling us to easily add at any moment additional components for image processing or text processing in order to improve the search results. Also, the results can be viewed in different formats according to user needs. For the next period we intend to add more components to this main architecture and to address the credibility component in the search process.

4. Participation in evaluation campaigns

Algorithms were tested already in the CLEF evaluation campaigns, like ImageCLEF (PlantIdentification), CHiC, QA4MRE and in SemEval (Sentiment Analysis in Twitter track).

Thus CEA group has worked on the Sentiment Analysis in Twitter track of SemEval and their run was ranked 5th out of 29 participant groups to the "**Task A: *Contextual Polarity Disambiguation*"** (www.cs.york.ac.uk/semeval-2013/task2/). With some adaptations, the method developed will be useful for credibility estimation in the project. Also, CEA group has participated to CLEF CHiC (Cultural Heritage in CLEF - http://www.promise-noe.eu/chic-2013/home) in order to evaluate multilingual implementation of ESA in two settings: ad-hoc retrieval and semantic enrichment. In the end they were ranked 2nd out of 7 participants for ad-hoc retrieval and 1st out of 2 for semantic enrichment.

The UAIC group was involved in the Plant Identification task at CLEF, where their group was ranked 5th out of 12 participating groups. Many parts from this system will be used in the architecture of the system which will be developed in MUCKE. Also, UAIC group participated in QA4MRE task at CLEF, but at this section the general results will be released at the CLEF conference in September.

5. Conclusions

We have shown in Section 2 the current status for tools and resources of all partners involved in the MUCKE project. In Section 3 we presented the skeleton of the system designed to be the internal sharing mechanism prototype. Section 4 presents how our algorithms were tested within existing relevant evaluation campaigns like CLEF and SemEval.