Computer Science Journal of Moldova, vol.**22**, no.3(66), 2014

Diversification in an image retrieval system based on text and image processing*

Adrian Iftene, Lenuta Alboaie

Abstract

In this paper we present an image retrieval system created within the research project MUCKE (Multimedia and User Credibility Knowledge Extraction), a CHIST-ERA research project where UAIC[1] is one of the partners[2]. Our discussion in this work will focus mainly on components that are part of our image retrieval system proposed in MUCKE, and we present the work done by the UAIC group. MUCKE incorporates modules for processing multimedia content in different modes and languages (like English, French, German and Romanian) and UAIC is responsible with text processing tasks (for Romanian and English). One of the problems addressed by our work is related to search results diversification. In order to solve this problem, we first process the user queries in both languages and secondly, we create clusters of similar images.

Keywords: image retrieval, search diversification, text processing and image processing.

1 Introduction

In the last years, the web has become a support for social media where users of social networks are pushing multimedia data directly from cameras, phones, etc. MUCKE project addresses this stream of multimedia

*This work was supported by MUCKE (Multimedia and User Credibility Knowledge Extraction) project Ref. No. 2 CHIST-ERA, 01.10.2012

[1]"Alexandru Ioan Cuza" University of Iasi

[2]Together with Technical University from Wienna, Austria, CEA-LIST Institute from Paris, France and BILKENT University from Ankara, Turkey

social data with new and reliable knowledge extraction models designed for multilingual and multimodal data shared on social networks. One of the aims of this project is related to give a high importance to the quality of the processed data, in order to protect the user from an avalanche of equally topically relevant data. In this context we built a novel image retrieval framework which performs a semantic interpretation of user queries and returns diversified and accurate results.

Over time, various theories involving search results diversification have been developed, theories that have been taken into consideration [1]: (i) *content* [2], i.e. how different are the results to each other, (ii) *novelty* [3], [4], i.e. what does the new result offer in addition to the previous ones, and (iii) *semantic coverage* [5], i.e. how well covered are the different interpretations of the user query. In the MUCKE project, we create a collection with over 80 millions images with their associated metadata, mainly extracted from Flickr[3]. Over this collection we perform processing at text level (on associated metadata) and at image level. These processing modules help the information retrieval system to retrieve images and to offer results in a diversification manner.

2 MUCKE Project

In the MUCKE project, for the moment, we have a collection of approximately 80 million images and their associated metadata that have been downloaded mainly from the Flickr database. Over this collection, we perform several processing tasks at both textual (on associated metadata) and image level and retrieve the results in a diversified way. Until in the end of project we intend to increase this collection to around 100 million images.

Over this collection, we built a system which will enable users to retrieve multimedia content. Figure 1 shows an overview of the MUCKE framework, covering how documents are processed, concepts extracted and indexed, similarity computed based on concepts, text and images, and how credibility is estimated and fed into the re-ranking process to

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

improve the final set of results.

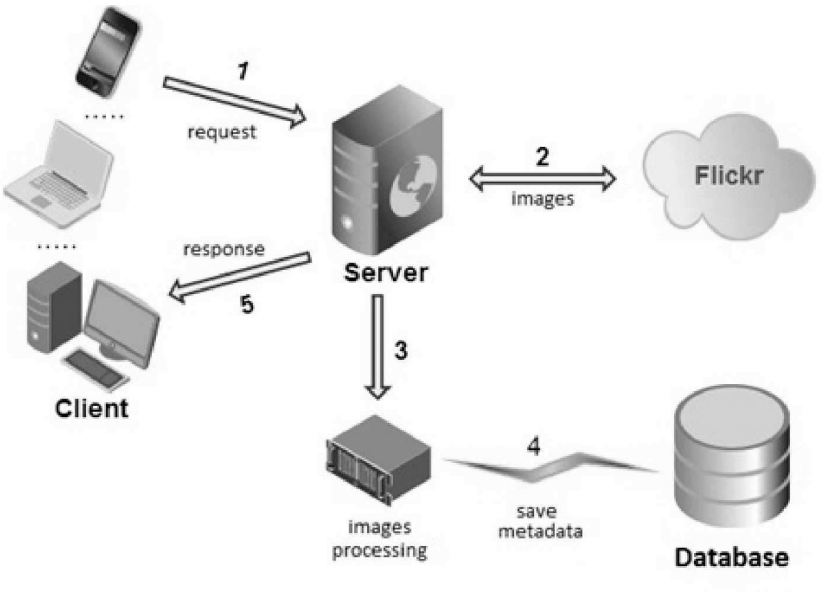

Figure 1. *MUCKE Architecture [6]*

3 Text processing module

The text processing module is used to process the associated metadata and to process the user query. For text processing, standard tools are used for POS-tagging, lemma identification and named entity identification. After text processing of associated metadata, the collection is indexed with Lucene. Then, the diversification is made on user query with Yago ontology and a special module for query expansion.

Yago ontology comprises the well known knowledge about the world [7]. It contains information extracted from Wikipedia and other sources like WorldNet and GeoNames and it is structured in elements called entities (persons, cities, etc.) and facts about these entities (which person worked in which domain, etc.). For example, with Yago

we are able to replace in a query like "tennis player on court", the entity "tennis player" with instances like "Roger Federer", "Rafael Nadal", etc. Thus, instead to perform a single search with initial query, we perform more searches with new queries, and in the end we combine the obtained partial results in a final result.
In our system we used Wikipedia and we created a resource based on categories that we found here. To accomplish this, we started with Romanian Wikipedia which has 8 groups of categories: culture, geography, history, mathematics, society, science, technology, privacy. In turn, these categories have subcategories or links to pages directly, as follows: Culture (30) (among which we mention photo, architecture, art, sports, tourism, etc.) Geography (15) (among which mention Romania, Africa, Europe Countries, maps, etc.), History (6) (among which mention After the recall, By region, etc.), Mathematics (11) (among which mention Algebra, Arithmetic, Economics, Geometry, Logic, etc.), Society (22) (among which mention Anthropology, Archaeology, Business, Communications, Philosophy, Politics, etc.), Science (23) (among which mention Anthropology, Archaeology, Astronomy, Biology, etc.), Technology (19) (among which mention Agriculture, Architecture, Biotechnology, Computer, etc.), Private life (8) (among which mention the Fireplace, Fun, People, Health, etc.). In the end, we obtained 8 big groups with 134 categories, which are subdivided into several subcategories and pages (hierarchical depth depends on each category and subcategory). In general, this hierarchy covers most of the concepts available for Romanian. For example, for Sport, we obtained 70 subcategories containing other subcategories and 9 pages. Going through these categories and subcategories, we built specific resources with words that signal concepts of type person, location and organization. Some examples of signal words from these categories are:

- For Person: "*acordeonist, actor, inginer, antropologist, arheolog, arhitect, femeie, arhivist, asasin, astronaut, astronom, astrofizician*, etc." (En: *accordionist, actor, engineer, anthropologist, archaeologist, architect, woman, archivist, assassin, astronaut, astronomer, astrophysicist.*) This is the biggest resource with over 391 signal words.

- For Location: "*continent, ţară, oraş, comună, sat, regiune, munte, râu, fluviu, piaţă, stradă, bulevard, târg, instituţie, universitate, spital, teatru,* etc." (En: *continent, country, city, township, village, region, mountain, river, market, street, avenue, fair, institution, University, hospital, theatre*).
- For Organization: "*companie, SRL, partid, grupare,* etc." (En: *company, LLC, party, group).*

Starting from a query that includes tennis players, it decides to use YAGO because our system identifies this word in the list with signal words for type person, and it calls a Sparql query with the following form:

```
PREFIX yago:<http://yago-knowledge.org/resource/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>

select ?instance ?category ?length where
{
       {select distinct ?instance
         where{
                ?class rdfs:label "tennis players"@ron.
                ?category rdfs:subClassOf ?class.
                ?instance rdf:type ?category.
         }
   LIMIT 5000
       } .
       ?instance yago:hasWikipediaArticleLength ?length.
       ?instance rdf:type ?category.
       ?class rdfs:label "tennis players"@ron.
       ?category rdfs:subClassOf ?class.
}
order by desc(?length) LIMIT 2000
```

After performing a search on Google with the word *"tennis players"* we obtain the results from Figure 2.

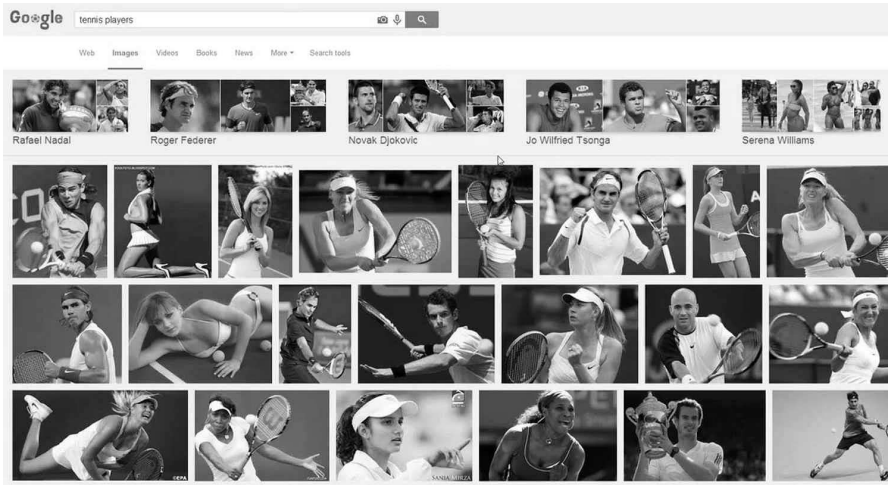

Figure 2. *Results offered by Google Image Search for the query "tennis players"*

After performing the same search in our application with the *"tennis players"* we obtain the results from Figure 3.

In our example and in other cases that we have studied, Google results are similar from the point of view of concepts presented in images returned. But, in the case of our application, there are more "colours" and more concepts in comparison with results offered by Google.

Query expansion module uses a technique of processing a given query in order to obtain new ones that are both more efficient and more relevant in the context of information retrieval operations. In this case, we faced with two major issues that occur when the end user enters the query: it is *not precise enough*, meaning that there are too many results returned, most of them being irrelevant or it is *not abstract enough*, meaning that the search does not return any results at all. Here, we applied two approaches: (1) *a global technique*, which analyzes the body of the query in order to discover word relationships (synonyms,

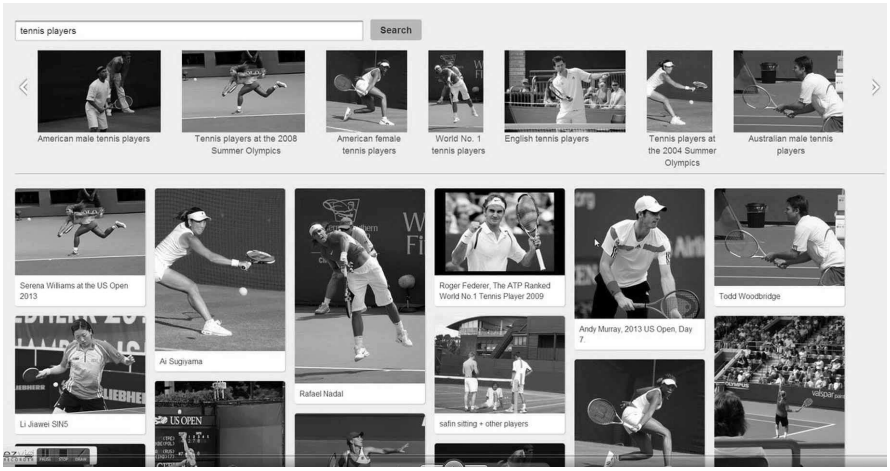

Figure 3. *Results offered by our application for the query "tennis players"*

homonyms or other morphological forms from WordNet[4]), to remove stop words (the, a, such, at) and to correct any spelling errors; (2) *local feedback* which implies the analysis of the results returned by the initial query, leading to re-weighting the terms of the query and relating it with entities and relationships originating from the target ontology.

4 Image processing module

The information retrieval system searches in its collection with the new obtained queries and returns a collection of relevant metadata with their associated images. The image processing module performs diversification on these returned results. The main aim of this module is to create clusters with similar images. Therefore the user will see only a representative image from every cluster and he/she is protected by various irrelevant images. In this way the similar images are hidden,

[4]WordNet: http://wordnet.princeton.edu/

the user will not be overwhelmed and will be able to see all the pictures on request.

For that, with the help of Mathlab and its predefined functions we extracted visual characteristics such as shape, color, texture, etc. Also, a naive algorithm that calculates an euclidean distance between the average color of the two images was implemented. Then, the clustering module organizes in clusters the images, according to certain metrics conducted from their features and based on the distances between them. We propose a density based algorithm inspired by DBSCAN [8], from which we take the ideas of *distance* and *dynamically creation of clusters*. The minimum number of images for the creation of a cluster is set to one, reason why, after clustering, isolated images (noise) cease to exist. Another difference concerns the input of the search, where the first image from the request is taken and not a random one, like in the case of DBSCAN.

In this approach, the notion of *distance* changes according to the type of request specified by the user: based on the *text* or on the *content*. In the first case, we use the Levenshtein algorithm [9] to calculate the distance between two strings, applied on the image's annotations (title, description, tags). In the second case, we compute the distance between images at the pixel level, by applying the algorithms provided by the AForge.NET[5] framework.

Using the algorithm described above we obtained the resulting clusters which represents the application's output.

5 Conclusions

In this paper we present our current work in MUCKE project. The paper addresses the diversification aspects that can be useful for an image retrieval system. For that we perform text processing on user query with Yago and WordNet resources and image processing in order to create clusters with similar images. The obtained results on the available collection are comparable and better in some perspectives

[5] AForge.NET: http://www.aforgenet.com/framework/

with other systems.

6 Acknowledgments

The research presented in this paper was funded by the project MUCKE, number 2 CHIST-ERA/01.10.2012.

References

[1] M. Drosou and A. Pitoura. *Search result diversification.* SIGMOD Rec., vol. 39, no. 1, New York, NY, USA, ACM (2010), pp. 41–47.

[2] S. Gollapudi and A. Sharma. *An axiomatic approach for result diversification.* Proceedings of the 18th international conference on World wide web (WWW), New York, NY, USA, ACM, (2009), pp. 381–390.

[3] J. G. Carbonell and J. Goldstein. *The use of mmr, diversity-based re-ranking for reordering documents and producing summaries.* SIGIR'98 Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM, (1998), pp. 335–336.

[4] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Bttcher, and I. MacKinnon. *Novelty and diversity in information retrieval evaluation.*, SIGIR '08 Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM, (2008), pp. 659–666.

[5] W. Zheng, X. Wang, H. Fang, and H. Cheng. *Coverage-based search result diversification.* Journal Information Retrieval, vol. 15, issue 5, Kluwer Academic Publishers Hingham, MA, USA, (2012), pp. 433–457.

[6] R. Bierig, C. Serban, A. Siriteanu, M. Lupu, and A. Hanbury. *A System Framework for Concept- and Credibility-Based Multimedia Retrieval.* In ICMR'2014 Proceedings of International Conference on Multimedia Retrieval, Glasgow, Scotland, April 2014, pp. 543–550.

[7] J. Hoffart, F. Suchanek, K. Berberich, and G. Weikum. *YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia.* Elsevier, Artificial Intelligence, vol. 194 (2013), pp. 28–61.

[8] M. Ester, H.P. Kriegel, J. Sander, and X. Xu. *A density-based algorithm for discovering clusters in large spatial databases with noise.* In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA. AAAI Press (1996), pp. 226–231.

[9] V. I. Levenshtein. *Binary codes capable of correcting deletions, insertions and reversals*, Soviet Physics Doklady, (1996), 707–710.

Adrian Iftene, Lenuta Alboaie Received September 19, 2014

Adrian Iftene
"Alexandru Ioan Cuza" University of Iasi, Faculty of Computer Science
Berthelot, 16, Iasi 700483, Romania
Phone: +40 232 201090
E–mail: *adiftene@info.uaic.ro*

Lenuta Alboaie
"Alexandru Ioan Cuza" University of Iasi, Faculty of Computer Science
Berthelot, 16, Iasi 700483, Romania
Phone: +40 232 201090
E–mail: *adria@info.uaic.ro*