Image annotation network
PFE report

Mael Ogier

INSA Lyon - Département Informatique mael.ogier@insa-lyon.fr

Abstract. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Keywords: object recognition, image transformations, local image features

1 Introduction

1.1 Context

This PFE was done in the context of the double master degree Informatique - Information und Kommunikation (IFIK), which brings together two Master programs : a degree in computer engineering at the National Institute of Applied Sciences in Lyon (INSA Lyon) and a Master in Informatik (Schwerpunkt : Information und Kommunikationssysteme) at the University of Passau.

I realised my project at the University of Passau during the last semester of my 5th year of study. During these 6 months, I had the amazing opportunity to discover the way of study in Germany and the organization of the University of Passau. There are 12,000 students in Passau (for 50,000 inhabitants) and **XXX** study computer science. The FIM faculty (Fakültät für Informatik und Mathematik) is composed of **XXX** chairs, I was part of Pr. Kosch's which currently employ **XXX** master students and **XXX** doctors students from different countries (one of the student I shared the office with was from Tunisia).

I worked on my own on this project under the supervision of two tutors, Dr. David Coquil on the german side and Dr. Elöd Egyed-Zsigmond on the french side. They were a precious help during the whole PFE process in term of advices and suggestions of work.

1.2 Background

Image is a popular medium nowadays : it is easy to capture, can be really light on your electronic device and speaks to everyone without distinction of language.

The huge production and consumption of images implies the need of an efficient way to store and search for the relevant one when the time comes. The best illustration to this need is to think of the nice but long moments one had with its relatives searching for the good picture of the newborn nephew in the family pictures album.

Since an image itself doesn't have a natural plain-text representation the best way to describe it is to add meta-data (data about the data) such as its date of creation, its dimensions or, and this is what this thesis is about, some tags.

There are a lot of ways if one wants to annotate pictures. We can do it manually, using our own words (like "Dad", "Home" ...), we can also analyze the raw picture, its pixel representation and compare some metrics (like the color histogram) to sample images in order to detect known concepts. Moreover, if the image already possesses annotations, we can enrich it semantically.

This field is so wide that it is impossible to speak about all the possibilities and technologies. In this study, we will focus on the last point and investigate the automation of the semantic enrichment. We will study the resources at our disposal and propose a solution keeping in mind the facts cited previously.

1.3 State of the Art

Semantic Web Linguistic semantics is the study of meaning that is used for understanding human expression through language. It is easy for two human-being to communicate (given that they speak the same language) and to understand what their partner say even if he's using a tricky turn of phrase. However, this task becomes way more difficult when it comes to the comprehension of the human language by a machine. How can the computer guess that "I am totally dead" means in fact "I am really tired" and that the speaker isn't actually dead ? Machines need structured resources to understand us and the Semantic Web is one of them.

The notion of "Semantic Web" has been mentioned for the first time by Berners-Lee et al in [1]. In this paper, they describe it as a Web which is readable by machines in opposite of most of Web's content which were designed for humans to read. The Semantic Web isn't a separate Web but an extension of the current one which will bring structure to the meaningful content of Web pages.

Two main technologies are used for the development of the Semantic Web : eXtensible Markup Language (short XML) and the Resource Description Framework (short RDF). XML allows everyone to create their own tags and to arbitrary structure their documents but gives no information about what this structure means. Meaning is provided by RDF which stores it in sets of triples which are composed of a subject, a predicate and an object. Those three components can be related to the subject, the verb and object of an elementary sentence. In [2], Miller presents a short introduction to the RDF standard and precises that a "Resource" can be any object which is uniquely identifiable by a Uniform Resource Identifier (URI).

The third basic component of the Semantic Web are collections of information called ontologies. An ontology is, in computer science, a document which defines the relations among concepts. Basically, Web ontologies are composed of a taxonomy, which defines classes of objects and their relations, and a set of inference rules.

Similarity Measures

DBpedia DBpedia[1] is a project originally launched by two German universities (Berlin and Leipzig) and backed by an important community. It explores Wikipedia[2] and extract information from it which results on the creation of a multilingual, large-scale knowledge base. The extraction framework, all the available end-points as well as some facts and figures about the project are presented in [3].

DBpedia's ontology is based on classes (320 items) which form a subsumption hierarchy, the root element being owl:Thing, with a maximal depth of 5[3]. Theses classes are described by a total of 1650 different properties, forming a large set of RDF triples (580 million extracted from the English version of Wikipedia).

[1] http://wiki.dbpedia.org/
[2] https://en.wikipedia.org/
[3] Complete classes tree : http://mappings.dbpedia.org/server/ontology/classes/

Even though DBpedia is now a worldwide project and provides pages in 125 languages, the English one is still the most represented. We can indeed find 4.58 million of things[4] including 1,445,000 instances of the class *Person*, 735,000 places *Place*, 251,000 *Species* ... The number of instances described in this language is about three time larger than the second and third language (French and German).

As well as any RDF-structured dataset, DBpedia can be requesting with SPARQL (which is an recursive acronym : SPARQL Protocol and RDF Query Language) queries. SPARQL allows the user to search, add, modify or delete RDF data available on the Internet, see [4] for more details about the language.

DBpedia also provides useful web services and HTTP endpoints. DBpedia Spotlight, which highlight DBpedia concepts in an input text is described in [5] and further details about disambiguation using this service are presented in subsection **??**. The official DBpedia SPARQL endpoint[5] allows the user to send SPARQL queries to the online Virtuoso Triple Store by using the browser interface or by sending a HTPP request. We learn in [3] that the average amount of hits per day of this endpoint is of 2,910,410 for the 3.8 dataset version.

WordNet WordNet[6] is a lexical database of English which has been presented for the first time in 1995 in [6]. It is hosted by the Princeton University, currently running version 3.1 but there are no current plan for a future release due to limited staffing.

Its structure is based on the concept of "synset" (synonym set), a set cognitive synonyms. WordNet distinguish among Types (common nouns, verbs...) and Instances (specific persons...). Synsets are interliked using conceptual, semantic and lexical relations. The hierarchy is built by the use of the super-subordinate relation (or hyperonymy, hyponymy in WordNet's jargon). These relations implements the two directions of the "IS-A" expression. For instance, *fruit* is **hyperonym** of *apple* and *horse* is a **hyponym** of *animal*, the root element being "entity". Other relations are also provided, like the antonymy (opposite of synonymy) or the meronymy and its opposite holonymy which implements the "IS-PART-OF" relation : *finger* is a **meronym** of *hand*. All these relations are transitive.

This resource is useful if we are searching for entities. Since the maximal depth is of the ontology is of 16, the leafs are very detailed nous (tsetse-fly, Yukon white birch, ...) but it also contains more general concepts (vehicle, animal, ...). WordNet contains at the moment 155,287 unique strings incuding 117,798 nouns.

It exists several ways to browse this resource. An online interface allows the user to manually query the dataset and to navigate in it through hyperlinks. For software and research purposes, the user has to download one of the released version of WordNet's dataset as well as a specific library according to the code language he's using.

[4] http://wiki.dbpedia.org/about/facts-figures
[5] http://dbpedia.org/sparql
[6] https://wordnet.princeton.edu/

Existing approaches

Mixing Statistics and Wordnet In [7], Jin and al. propose the integration of the WordNet (1.3) semantic resource in a statistic-based annotation process in order to remove irrelevant keywords.

Their algorithm is organized as follows : they first generate a set of keywords with the help of a statistical model called *Translation Model* (TM). Some of those candidates tags are relevant and some aren't.
In order to filter the irrelevant ones, they then compute several semantic similarity measures (Lin, Jiang and Conrath and Banerjee and Pedersen). Finally, they combine these metrics using Dempster-Shafer Theory ([8]) and the keywords with a resulting score under a chosen threshold are removed. Detailed method steps are presented in the original paper.

Concerning their results, they compare their *TMHD* proposed approach with a basic TM process. Based on a set of most frequently used keywords, they found that, on average, precisions values of TM and TMHD are respectively 14.21% and 33.11%. This indicates that TMHD is 56.87% better than TM. It is interesting to note that the recall score stays the same due to the fact that only irrelevant keywords are removed. They also compare *TMHD* to the use of individual measures with TM and the results aren't as good as those from their combination.
These results show the power of knowledge-based data and similarity measures when it's added to a statistical model.

Graph-cut based enrichment In [9], Qian and Hua expose their graph-based approach of the tag enrichment process. They represent each initial tag of their corpus as a node and interlink them (using *n-links*). The weight of those n-links can be seen as the similarity between the two linked nodes, computed by the help of the Google distance [10]. They add two virtual nodes called *sink* and *source*. Then, they link all nodes to one of these virtual nodes using *t-links*.

The aim of their approach is to split all the tags into two distinct sets S (containing the source node) and T (containing the sink one) by assigning the labels s (source) if the tag is relevant to the image and t (sink) if not to the nodes. Then, they determine how many tags are relevant to the image by solving the combinational optimization problem through the graph.

This paper is really short and not very clear but it gives good ideas about the tags' representation as a graph and how to interlink them. According to the authors, the results are satisfactory.

Enrich Folksonomy Tag Space Folksonomies are typical Web 2.0 systems that allow users to upload, tag and share content such as pictures, bookmarks ... In [11], Angeletou and al. envisaged tag space enrichment with semantic relations by exploring online ontologies. Their method is composed of two phases :

- Concept identification
- Relation discovery

The first step is achieved by extracting concepts from online ontologies in which the local concept label matches the tag. In order to exploit all meanings, the authors retrieve all the potential semantic terms for each tag and then discover relation between them in the second phase. This

means that no disambiguation is processed but it is a consequence of the relation discovery phase. This phase consist of the identification of the relation between two tags *T1* and *T2*. Four kind of relations are distinguished : Subsumption, Disjointness, Generic, Sibling and Instance Of. These relations can be found by two ways : a relation can be declared within an ontology or, if no ontology contain such relation, one is made by crossing knowledge from different ontologies.

The author then present different experiments as well as some issues rose during this phase. One in particular is important to keep in mind : when users tags resources, especially pictures, they tend to tag them with specific vocabulary, mainly instances rather than *abstract* concepts. This can result on lot of "semantic noise" : tags which can't be match with concepts from online ontologies.

This paper is really interesting and approaches the topic in a very general point of view, which ensure the flexibility of its implementation. We will see in 1.4 that this method perfectly adapt to our study.

1.4 Method

With this work we want to propose a prototype which semantically enrich images given an initial set of tags. This prototype will be based on 3 steps :

1. Concept identification
2. Relation discovery
3. Candidates detection

As previously said, the two first steps will be similar as those presented in 1.3. The difference we want to propose is to use several online ontologies in order to detect concepts and to create relations between them. We selected two resources : DBpedia and WordNet, already presented above.
The last step of our method is the detection of potential new tags by using the graphs previously created. Three experiments are proposed based on this. We also further investigated DBpedia's environment by directly using its sources (the Wikipedia's pages) for two other experiments.

2 Contribution

2.1 Architecture

In order to support the method presented in 1.4 some technology choices were made. In this section we will present the most important of them and especially detail the structure of our graph-based data model.

Technologies

Java Language Java is a general-purpose computer programming language that is concurrent, class-based, object-oriented, and specifically designed to have as few implementation dependencies as possible. It is also a cross-platform language which means that it would be possible to use it without any recompilation needed. Java is very easy to use, well documented a has the support of a large community (more than 9 million developers reported). Therefore, lot of libraries are available, we will present some of them below.

Neo4j Graph-based databases are very intuitive to work with and allow the user to model the world as he experience it. The model schema isn't rigid and the user can edit it at anytime, adding new entities or new kind of relationships. Neo4j is an open-source graph database, implemented in Java (so cross-platform), maturing for 15 years and currently running version 2.2. It is the most popular graph database nowadays[7], has a great scalability, a strong community and has its own query language : Cypher.

Semantic Resources Libraries We needed to access the chosen online ontologies (DBpedia's and WordNet's) from our prototype. To achieve this, we used the fact that Java is very popular and lot of libraries are available.
We chose Apache JENA ARQ[8] to query the RDF-base schema of DBpedia. This solution is stable and maintained by a famous structure : Apache. Using it was really simple.
Regarding WordNet, we used JAWS[9] which has been developed and is maintain by a member of the Southern Methodist University (Dallas, Texas). Its last version is a bit old but this isn't an issue since WordNet's upgrades have also stopped. This library was also deeply intuitive and easy to use.

JSoup Our two last experiments are based on Wikipedia's web-pages. Therefore we needed a way to crawl and extract content from them. The JSoup[10] library was a perfect asset to achieve this. It is open-source, implements the WHATWG HTML5 specification, and parses HTML to the same DOM as modern browsers do. It also allows the user to build specific queries to access particular elements in the DOM.

[7] http://db-engines.com/en/ranking/graph+dbms
[8] https://jena.apache.org/documentation/query/index.html
[9] http://lyle.smu.edu/ tspell/jaws/
[10] http://jsoup.org/

Stanford NLP Crawling web-pages is a thing, but extracting relevant data from it is another one. The Stanford NLP Research Group[11] has released several libraries in different programming languages including Java. Those libraries can achieve many things such as sentence segmentation, Part-of-speech (POS) tagging, named entities recognition and so on... We used the POS Tagger to extract nouns from Wikipedia paragraphs.

Graph Structure

Vertexes

Edges

Pro-Cons

2.2 Experiments

Implementation Explanation

Results and Analysis

3 Perspectives

Add new resources Work with other similarity distances

[11] http://nlp.stanford.edu/

4 Conclusion

References

1. Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic web." Scientific american 284.5 (2001): 28-37.
2. Miller, Eric. "An introduction to the resource description framework." Bulletin of the American Society for Information Science and Technology 25.1 (1998): 15-19.
3. Lehmann, Jens, et al. "DBpedia-a large-scale, multilingual knowledge base extracted from wikipedia." Semantic Web Journal 5 (2014): 1-29.
4. Prud, Eric, and Andy Seaborne. "Sparql query language for rdf." (2006).
5. Mendes, Pablo N., et al. "DBpedia spotlight: shedding light on the web of documents." Proceedings of the 7th International Conference on Semantic Systems. ACM, 2011.
6. Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.
7. Jin, Yohan, et al. "Image annotations by combining multiple evidence & wordnet." Proceedings of the 13th annual ACM international conference on Multimedia. ACM, 2005.
8. Shafer, Glenn. A mathematical theory of evidence. Vol. 1. Princeton: Princeton university press, 1976.
9. Qian, Xueming, and Xian-Sheng Hua. "Graph-cut based tag enrichment." Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011.
10. Cilibrasi, Rudi L., and Paul Vitanyi. "The google similarity distance." Knowledge and Data Engineering, IEEE Transactions on 19.3 (2007): 370-383.
11. Angeletou, Sofia, et al. "Bridging the gap between folksonomies and the semantic web: An experience report." (2007).