Classifying Image Galleries into a Taxonomy Using Metadata and Wikipedia

Gerwin Kramer[1], Gosse Bouma[1], Dennis Hendriksen[2], and Mathijs Homminga[2]

[1] Information Science, University of Groningen
gerwinkramer@gmail.com, g.bouma@rug.nl
[2] Kalooga, Groningen
{dennis.hendriksen,mathijs.homminga}@kalooga.com
http://www.kalooga.com

Abstract. This paper presents a method for the hierarchical classification of image galleries into a taxonomy. The proposed method links textual gallery metadata to Wikipedia pages and categories. Entity extraction from metadata, entity ranking, and selection of categories is based on Wikipedia and does not require labeled training data. The resulting system performs well above a random baseline, and achieves a (micro-averaged) F-score of 0.59 on the 9 top categories of the taxonomy and 0.40 when using all 57 categories.

Keywords: gallery, image gallery, classification, hierarchical classification, taxonomy, Wikipedia.

1 Introduction

Organizing and managing the overwhelming amount of images on the web has become an active area of research. Image classification is the task of (automatically) classifying images into semantic categories. For this purpose, either visual clues can be used or text and metadata surrounding an image.

In this paper, we concentrate on a special case of image classification, namely *image gallery*[1] classification. An image gallery is a (part of a) website that displays a collection of images. Usually, this collection has a certain topic in common like a person, group, event or place. We are especially interested in these galleries, because they are often suitable as illustration in (on-line) news items. By categorizing galleries, we hope to improve the accuracy of a method that finds suitable illustrations for news items. We only use textual metadata, such as the title and image captions for classification. Galleries are classified into a hierarchically structured set of predetermined categories, our taxonomy.

The traditional, statistical, supervised, approach to image classification requires an (up-to-date) labeled training-set, which can be hard to create and

[1] Example galleries can be extracted from several locations on the web, e.g. IMDB (http://www.imdb.com), Flickr (http://www.flickr.com/), and various publisher's websites.

G. Bouma et al. (Eds.): NLDB 2012, LNCS 7337, pp. 191–196, 2012.

maintain. We avoid the use of a training-set by adopting an ontology-based classification technique based on [3]. The ontology we use is extracted from the category system of Wikipedia. Entities found in the metadata are linked to Wikipedia pages, and the most important Wikipedia categories of those entities are chosen as categories for the image gallery.

2 Related Work

The idea of creating a taxonomy of image galleries is not totally new. ImageNet [1] aims to (manually) assign images to concepts in WordNet. The goal is to provide a visual data set to be used for training and benchmarking classification algorithms. Text-based approaches to image classification can be tracked back to 1970s [15]. An ontology can be beneficial for text-based classification, for instance to bridge 'lexical' or 'semantic gaps' [11,10].

For the purpose of ontology-based research, the online and open encyclopedia Wikipedia is a valuable source [4]. A Wikifier [6,7,2] is a system that links entities to Wikipedia pages. Wikipedia can be used to calculate *relatedness* between terms as well [12]. The relatedness measure can be used to rank concepts in a text by importance (topic indexing) [5,9]. Most articles in Wikipedia are assigned one or more, hierarchically structured, categories. [14] acknowledge that the Wikipedia category graph can be used for various NLP tasks, although using it as is, has several drawbacks. [8] extract a taxonomy from the Wikipedia category graph that is more suitable for NLP applications.

3 Taxonomy

Our taxonomy is a hierarchical structured set of categories for image classification. Its categories are chosen by their relevance for image gallery classification and the needs of the Kalooga application that automatically suggests galleries for news items.

The root category of the taxonomy is *Contents*. Image galleries that fit none of the descendant categories should be classified here. The first-level categories are *People* (has 11 descendant categories and 2 sublevels), *Sports* (has 28 descendant categories and 2 sublevels), *Vehicles* (has no descendant categories), *Places* (has 2 descendant categories and 2 sublevels), *Entertainment* (has 8 descendant categories and 2 sublevels), *Arts* (has no descendant categories), *Animals* (has 6 descendant categories and 1 sublevel), and *Plants* (has no descendant categories).

Each category in the taxonomy is linked to a Wikipedia category from the Wikipedia category graph. This mapping enables us to perform the entity classification process described in section 4. Categories could in theory be mapped to a category graph in another language as well to enable classification of galleries in that language.

In hierarchical classification, a gallery should be classified in the most specific category or categories. If no category is suitable, *Contents* should be assigned. The decision was made that a category should only be assigned when it applies to each of the images of the gallery.

4 Methodology

The intuition behind the presented method is that shared and/or relevant categories of important entities in gallery text will be categories of the gallery. In the following subsections, each step of this method is described.

Extract Gallery Text. Metadata, such as the URL, title, description, image captions, and keywords found in the HTML of the gallery is merged into one text (URL's are decomposed into words). The title and URL words are included twice, to boost their weighting. Stopwords (including typical 'gallery' words such as *picture* and *gallery*) are removed.

Extract Entities. For the process of recognizing entities from the text, an in-house *wikifier* is used that recognizes entities and links these to Wikipedia pages. A semantic graph is constructed in which the nodes are entities and the edges the semantic relatedness between them. The computationally inexpensive Wikipedia link-based measure (WLM) is used [12] to compute relatedness. To filter out unlikely edges in the graph, relatedness must be above a certain threshold.

Score Entities by Importance. We use a variant of Averaged PageRank weighting (APW) [9], which takes into account the centrality of the entity in the semantic graph and the relative frequency of the entity.

For computing centrality, we use a variant of *closeness centrality* [13]. In 1, we first calculate closeness centrality (CC) of an entity within its subgraph and then multiply by the size of the subgraph ($distance(a, b) = 1 - SR_{a,b}$ and g the size of the subgraph), because an entity from a large subgraph is intuitively of higher importance:

$$CCe_i = g \cdot \frac{g - 1}{\sum_j distance(e_i, e_j)} \tag{1}$$

The relative frequency of an entity is calculated using $tf\text{-}idf$, in which entities are treated as terms and Wikipedia as the document collection. The term count in the corpus is then the inlink count of the entity's article. The calculation of the total combined entity score is presented in (2). If the maximum centrality score is 0, then only the tf-idf part is used.

$$ESe, g = \frac{1}{2}\left(\frac{CCe}{CCmax} + \frac{tf\text{-}idfe, g}{tf\text{-}idfmax}\right) \tag{2}$$

After the scoring is applied to each entity, a threshold is applied to filter out noise.

Find and Score Entity Categories. In this step, we try to find taxonomy categories for each entity by using its Wikipedia article. Each article is connected to categories of the Wikipedia category graph. By searching through the ancestors of the article's categories in this graph, we try to find broader categories that are *also present in the taxonomy*. To prevent inclusion of highly general

or irrelevant categories, a maximum search depth is applied of 5 steps. Every retrieved category that matches a taxonomy category receives a relevance score.

$$ECS_{c,e} = \min(0.5, \frac{1}{d_{min}}) + \min(0.5, \frac{n_{path}}{{d_{avg}}^2}) \quad (3)$$

This entity's category score $ECS_{c,e}$ is the combination of the minimum distance d_{min} to travel from entity e to category c and the number of paths n_{path} from e to c, normalized by the average path distance d_{avg}. We like to classify the entity as specific in the taxonomy as possible. Therefore, if a certain category from the taxonomy is found, candidate ancestor categories are deleted.

Select Gallery Categories. The final step of the process is selecting categories for the gallery. For each category, we calculate the final gallery category score (GCS) as denoted in (4).

$$GCS_{c,g} = \sum_{k} ES_{e_k,g} \cdot ECS_{c,e_k} \quad (4)$$

The score is based on the amount of entities belonging to a candidate category c of gallery g, the importance of these entities (ES), and how relevant the category is to each entity (ECS). After calculating GCS for each candidate, a cut-off is applied to preserve only the most relevant categories.

5 Experiments

Test Set. Six persons manually assigned categories from the taxonomy to galleries. This resulted in a total of 734 English galleries of which 223 were randomly selected and 511 were found by manually finding galleries for categories. Most galleries (90%) had only one category.

Performance Measures. Because the number of galleries per category is very unevenly distributed, we report both micro and macro averaged precision, recall, and F-score.

In classic precision and recall measures, the hierarchical structure of the taxonomy is not taken into account. For hierarchical classification, this may be too pessimistic. Therefore, we measured precision and recall at the 9 top-level categories, and for all 57 categories.

Baseline. A random baseline classifier was created with 90 percent of chance to pick one category for a gallery and 10 percent chance to pick two categories for a gallery. The average of 100 runs is taken. A most frequent baseline is not used, because the distribution in the test set does not reflect the actual distribution of categories.

Thresholds. During tests on the development set, we found the following thresholds to result in a reasonable balance between precision and recall: Minimum semantic relatedness SR: 0.20; Minimum entity score ES: 0.45; Minimum entity category score ECS: 0.45; The gallery category score $GCS_{c,g}$ must be at least 50% of the highest GCS; Maximum number of gallery categories: 3;

6 Results

In Table 1, we can see that the classifier performs significantly better than the random baseline. The poor results of the random baseline give an indication of the complexity of the classification problem. Unsurprisingly, the classifier performs better in the 9 broader categories than for all 57 categories.

Table 1. Overall performance

Categories	Measure	Method	Precision	Recall	F-score
9 top-level categories	micro-avg	random	0.20	0.20	0.20
		classifier	0.54	0.66	0.59
	macro-avg	random	0.14	0.14	0.14
		classifier	0.54	0.70	0.61
All 57 categories	micro-avg	random	0.02	0.02	0.02
		classifier	0.35	0.46	0.39
	macro-avg	random	0.02	0.02	0.02
		classifier	0.48	0.47	0.48

The higher performance of the macro-averaged measures, compared to the micro-averaged measures indicates that the classifier performs relatively well on galleries from some small categories from the test-set and/or it performs relatively poor on galleries from some large categories from the test set. The lowest performance was achieved in the *Places* categories. This is largely due to most galleries being related to a place and entities on Wikipedia are highly related to places.

7 Conclusions and Future Work

We have seen how, and to what extent, image galleries can be classified into a taxonomy using metadata and Wikipedia. For the classification in large distinct categories, the presented method looks promising. However, fine grained classification is erroneous.

In future research, existing issues can be addressed by investigating some unexplored techniques. First, a change in the taxonomy might help to address low precision of the *Places* categories by making the location of a gallery a property. Second, the entity classification errors could be addressed by investigating additional ontologies and/or knowledge sources besides the Wikipedia category graph. Third, taking gallery context information into account might increase performance.

Finally, it would be interesting to find the *inter-annotator agreement* (IAA) about categories for gallery text. With the results of this test, a realistic target precision and recall for automated classification can be established. A comparison with a state of the art statistical classifier would also be interesting.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR 2009 (2009)
2. Hoffart, J., Yosef, M., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: Proc. of EMNLP, pp. 27–31 (2011)
3. Janik, M., Kochut, K.: Training-less ontology-based text categorization. In: ECIR Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2008), pp. 3–17. Citeseer (2008)
4. Medelyan, O., Milne, D., Legg, C., Witten, I.: Mining meaning from wikipedia. International Journal of Human-Computer Studies 67(9), 716–754 (2009)
5. Medelyan, O., Witten, I., Milne, D.: Topic indexing with wikipedia. In: Proceedings of the AAAI WikiAI Workshop (2008)
6. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: CIKM 2007: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, pp. 233–242. ACM, New York (2007)
7. Milne, D., Witten, I.: Learning to link with wikipedia. In: Proceeding of the 17th ACM Conference on Information and Knowledge Management, pp. 509–518. ACM (2008)
8. Ponzetto, S., Strube, M.: Deriving a large-scale taxonomy from wikipedia. In: AAAI, pp. 1440–1445. AAAI Press (2007)
9. Tsatsaronis, G., Varlamis, I., Nørvåg, K.: Semanticrank: ranking keywords and sentences using semantic graphs. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1074–1082. Association for Computational Linguistics (2010)
10. Wang, H., Liu, S., Chia, L.: Does ontology help in image retrieval?: a comparison between keyword, text ontology and multi-modality ontology approaches. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 109–112. ACM (2006)
11. Wang, P., Hu, J., Zeng, H., Chen, Z.: Using wikipedia knowledge to improve text classification. Knowledge and Information Systems 19(3), 265–281 (2009)
12. Witten, I., Milne, D.: An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In: Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, pp. 25–30. AAAI Press, Chicago (2008)
13. Wolfe, A.: Social network analysis: Methods and applications. American Ethnologist 24(1), 219–220 (1997)
14. Zesch, T., Gurevych, I.: Analysis of the wikipedia category graph for NLP applications. In: Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing, pp. 1–8. Association for Computational Linguistics, Rochester (2007)
15. Zhu, Q., Lin, L., Shyu, M., Liu, D.: Utilizing context information to enhance content-based image classification. International Journal of Multimedia Data Engineering and Management (IJMDEM) 2(3), 34–51 (2011)