Automatische Annotation von Bildern mit Assoziationsregeln

Automatic Image Annotation by Association Rules

Thorsten Hermes, Arne Jacobs
Center for Computing Technologies
University Bremen, Germany

E-mail: {hermes|jarne}@tzi.de

Adalbert F.X. Wilhelm
Jacobs University
Bremen, Germany

E-mail: a.wilhelm@jacobs-university.de

Zusammenfassung:

In dieser Arbeit beschreiben wir unseren Ansatz zur automatischen Annotation von Bildern basierend auf Assoziationsregeln. Die Regeln werden aus Artikeltexten und den damit verbundenen Bildern von Nachrichtenseiten im WWW gewonnen. Der Ansatz vereint drei unterschiedliche Suchphilosophien: Text-basierte Suche, reine visuelle Suche und domänenspezifische, bildverstehende Ansätze. Als einen ersten Schritt durch unser System zeigen wir, wie einzelne Webseiten quasi geerntet werden können, um sie anschließend partiell zu parsen und durch computerlinguistische Werkzeuge zu analysieren. Diese Analyseergebnisse werden dann untersucht und Assoziationsregeln bestimmt. Im Rahmen eines Experiments zeigen wir, dass Korrelationen zwischen visueller und textueller Ähnlichkeit existieren. Hierzu reichen reine syntaktische Bildähnlichkeiten basierend auf so genannten „low-level" Merkmalen aus.

Abstract:

In this paper we propose an approach to automatic image annotation based on association rule mining on article text and linked images from structured news websites. The approach tries to combine the advantages of three different image search and retrieval philosophies: Text-based search, pure visual search, and domain-specific image understanding approaches. As a first step towards the implementation of our framework, we show how selected websites can be automatically harvested, structurally parsed and processed by tools from Computational Linguistics, and how the results can be explored with association rule mining algorithms. In an experiment on the correlation between textual and visual similarity we show that such a correlation indeed exists, even with a purely syntactical image similarity measure based solely on low-level image features.

1. Introduction

The ever-increasing amount of image data on the World Wide Web requires effective automatic search and retrieval methods to facilitate access to and (re-)use of these images. Several approaches have been proposed to tackle the problem. One prevalent approach is used by most current WWW search engines. It tries to associate images with surrounding text or use meta-tags and filenames to provide textual queries on the image pool. Another technique, based on image processing, is to incorporate syntactic image features like color, and texture along with corresponding similarity measures to enable visual queries [3]. Further approaches, settled in the field of image understanding, try to recognize a fixed set of scenes or objects in a restricted domain to enrich images with searchable metadata. Successful examples have been shown in the domain of natural sceneries [6] and for scale-invariant detection of certain rigid objects, e.g., cars and motorcycles [4]. While the first approach requires an image to be associated with text to be searchable, the visual similarity approach can be applied to any image. However, the latter

remains completely syntactical, which limits its usefulness in semantic queries. The image understanding approaches shine here, but are by nature limited to very narrow domains.

We will present our approach to automatic image annotation, search, and retrieval, which combines the three different philosophies into a one framework. Therefore, it consists of three main parts: First, a large database of images associated with structured text on different semantic levels. This database is gathered by harvesting several popular news websites that feature strongly structured articles that can be parsed automatically. Each image will be associated with text on different levels, e.g., its caption, the surrounding article text, the corresponding article's title, and the domain of the article (e.g., politics, sports, financial news, etc.). By using natural language processing, keywords, named entities, etc. will be extracted on each level. Second, on an additional level, each image is associated with its syntactical image features like color and texture. Furthermore, local image descriptors that have been proven useful in the task of image understanding [4], will be extracted from and associated with each image. Third, this offers the possibility to link text/keywords of an article to the corresponding images in the same article, to link different images by their syntactical features and local descriptor, and even local descriptors in the same image. This huge amount of relations will be exploited by data mining techniques, specifically association rules, to extract hidden patterns or implicit knowledge.

Our first steps towards this approach are shown in this paper. In sections 2 and 3 we illustrate our news website harvesting, parsing and natural language processing processes. Section 4 addresses the association rule mining stage, followed by an introduction of our visual similarity features in section 5. First results on the correlation of textual and visual similarity between news articles are shown in section 6. We conclude with section 7.

2. News websites harvesting

Current popular news websites offer professionally edited content in a relatively structured form. We chose a small set of popular German news websites to be archived on a regular basis. Currently, we are using articles from www.tagesschau.de and www.faz.de, as these are sufficiently structured for automatic processing and we have implemented corresponding structural parsers. We furthermore collect articles from several other sites, which are not parsed at the moment. The raw data from the articles (mostly HTML and images) is regularly downloaded from these sites, and, in the case of the two mentioned sites, its structure is parsed, and the resulting text is inserted into a relational database. This is a fully automatic process. We incorporate a simple data model, consisting of an article data type including category (e.g., sports or politics), title, summary, and article text, and an image data type consisting of an image caption, a reference to the parent article, and the actual image data.

The database currently contains over 30,000 articles collected over several months. After harvesting and structural parsing, we use natural language processing techniques on the structured data as illustrated in the next section.

3. Natural language processing

Simple text processing, e.g., searching for text by exact matching etc., has many shortcomings. Ambiguity, synonyms, derivation, and flection make it difficult to work directly on unprocessed text data. Furthermore, very frequent words often add little information (so-called stop words). This can be overcome by filtering out stop words, based on pre-compiled lists, prior to further processing. Flection and derivation can to some extent be addressed by so-called stemming algorithms, like the popular *Porter stemmer* [7]. Stemming algorithms, however, remain purely morphological. We address stop words and flection by using a part-of-speech (POS) tagger, namely the *TreeTagger* developed by Schmid [8] and [9], which achieves an accuracy of ~97% on German text. It features 11 main part-of-speech tags, e.g. nouns, verbs, articles, etc., and 48 part-of-speech tags including subclasses. These subclasses contain further part-of-speech specific information, e.g., if a noun

refers to an entity name, or if a verb occurred in its infinitive form in the text. The TreeTagger also takes care of the tokenization of a given text.

In this first approach we consider only the main tags to filter out all words that are not nouns. This is because we regard the nouns as the words containing the information the most relevant to us. Nevertheless, the remaining results of the POS tagger are not discarded but kept for future processing. All results are stored in the database in a reverse index, which facilitates an efficient search for each word or combination of words in all texts. The nouns are used in the association rule mining stage, which will be discussed in the following section.

4. Association rule mining

Association Rules are a typical data mining procedure and aim to describe relationships between items that occur together. They are in particular suited to extract implicit knowledge from the data base and to make this knowledge accessible for further quantitative analysis. They have been proposed by Agrawal *et al.* [1] in the context of market basket analysis to provide an automated process, which could find connections among items, that were not known before. Discovery of association rules is based on the frequent item set approach. Typically some variation of the a priori algorithm [1] is used with the aim of generating all association rules that pass some user-specified thresholds for support and confidence. The problem is, that depending on the specified thresholds for confidence and support a vast amount of rules may be generated [2]. By restricting the association rule mining to specific domains, e.g., *sports* and/or *politics*, we will try to find the interesting rules for that domain.

As the possible article categories of the chosen news websites (*Tagesschau* and *FAZ*, namely) are not identical, we create a manual mapping between them to collect all articles of a category we wan to investigate. We chose the general categories sports and politics for our association rule mining. For the category sports we incorporate all articles from *Tagesschau* and *FAZ* that have the German word "Sport" in their website-specific category (this captures the sports articles of both websites). For the category politics we incorporate all articles from *FAZ* that contain the German word "Politik" in the FAZ-specific category, in addition to all articles from *Tagesschau* with the literal categories "Ausland" (i.e., international politics) and "Inland" (i.e., national politics).

The *item sets* we take as input to the association rule mining algorithm are sets of words (only nouns in our case) that occur in an article of the respective category. Because the number of different words that may occur in an article is still very high regarding only nouns (in the multiple thousands), we choose a subset of words in each of the two categories, based on the *tf-idf* (term frequency-inverse document frequency) measure. We do this to increase the relevance of the generated rules and at the same time reduce the computational complexity of the association rule mining process. We use the 250 nouns with the highest tf-idf measure for each category.

For the actual association rule learning we use the implementation of the *Apriori* algorithm [2] in the *Weka* machine learning tool [10]. We generate the 100 best rules based on the *lift* criterion with the maximum confidence parameter set to 0.2 (20%) and the minimum confidence parameter set to 0.005 (0.5%). Of these generated rules, we manually choose a subset for our experiments on the correlation of textual and visual similarity between news articles. The algorithm for the computation of visual similarity between two articles will be described in the next section.

5. Visual similarity

To compute the visual similarity between two given articles we first compute the visual similarity between each pair of images from both articles (one image from each of the two different articles). We then define the visual similarity between the two articles as the highest visual similarity among those image pairs. The algorithm is a derivative of the approach implemented in the PictureFinder system for graphical image search [3].

The visual similarity between two given images is based on the spatial distribution of localized color and texture features. Color features are computed based on the CIE Lab color space with a bias to the a and b plane. Texture features are based on the amount of edge points. Therefore, this measure implements a mapping to a qualitative texture energy measure a so-called visual property of texture [11] from homogeneous to non-homogeneous. Whereby on the one side a low amount of edge points in a certain region stands for homogeneous and on the other side a high amount of edge points in a certain region stands for non-homogeneous.

These values form a feature vector or a so-called signature of every image. A non-metric distance function in addition with the use of a heuristic concerning the position in the a and b plane serves as a similarity measure.

6. Experimental results

From the sports article set we manually choose the (automatically generated) association rule (*"Tour" & "de" & "Radsport" => "Doping" & "France"*) as an interesting result for our following experiment. The term "de" was classified as a noun by the POS tagger, as it is part of the string "Tour de France", which is classified as a noun. The German term "Radsport" translates to "cycling" in English. The above rule has a support of 136 articles in the sports category, and a confidence of 71% (96 of 136 articles containing "Tour" & "de" & "Radsport" also contain "Doping" & "France"). The 96 articles likely correspond to the coverage of one important sports news item during our news website harvesting, namely the doping scandals associated with the Tour de France 2007. The articles obeying the rule will be the base for our experiments on the correlation between textual and visual similarity.

From the politics article set we choose the association rule (*"Afghanistan" & "Kabul" => "Taliban"*), which corresponds to the coverage of German international politics in Afghanistan, where a small German military unit is still based at. The rule has a support of 166 articles in the politics category, with a confidence of 75% (125 of 166 articles).

For our experiment, for each of the articles that correspond to our chosen rule we compute the visual similarity between that article and all other articles in the respective category (sports for the first rule, politics for the second rule). For each article we get a ranked list of all articles in the respective category based on their visual similarity to that article. We compute the *average precision* for each article and average them, resulting in the *mean average precision* (MAP) for a chosen rule.

For the sports article set (which contains 2529 articles) we would expect a MAP value of about 3.8% for a random ranking of articles. Using the visual search algorithm as explained above we achieve a MAP value of 7.4%.

For the politics article set (which contains 4633 articles) we would expect a MAP value of about 2.7% for a random ranking of articles. Using our visual search algorithm we achieve a MAP value of 5.7%.

7. Conclusion and future work

We proposed a framework for automatic image annotation based on association rule mining on linked text and image data collected from structured news websites. As a first step, we performed an experiment on the correlation of textual and visual similarity, using a purely syntactical image similarity algorithm. The results show that there is such a correlation. As the mined association rules on the text data represent very high-level topics (i.e., the doping scandals of the Tour de France and the German military units based in Afghanistan), and the visual search algorithms use very low-level image features (i.e., color and texture), this correlation is not very high. We hope this can be improved by usage of intermediate-level features like the interest point detectors and corresponding local image descriptors.

Next steps are the application of interest point detectors and local image descriptors to supplement the currently color- and texture-based visual similarity measure. These descriptors could also be incorporated into the association rule mining, as they represent a kind of *visual word*. Association rule mining could then be used to find correspondences between (most likely only simple) concepts found in the article text and their visual representation in the articles' images. This might work specifically for named entities. A specialization of this approach with the focus on persons could lead to a fully automatically trainable face recognizer, which learns the persons' names from the article texts and which learns the corresponding face model from the associated images.

References

[1] Agrawal R., Imielinski T. and Swami A. (1993) – Mining associations between sets of items in massive databases. In *Proceedings of the ACM-SIGMOD 1993 International Conference on Management of Data*, Washington D.C., pp. 207–216.

[2] Agrawal R. and Srikant R. (1994) – Fast algorithms for mining association rules, Technical Report RJ9839, IBM, IBM Research Report RJ9839.

[3] Hermes T., Miene A. and Herzog O. (2005) – Graphical search for images by PictureFinder. In *Int. J. Multimedia Tools and Applications. Special Issue on Multimedia Retrieval Algorithmics*, 27(2), pp. 229–250.

[4] Leibe L. and Schiele B. (2004) – Scale-invariant object categorization using a scale-adaptive Mean-Shift search. In *C.E. Rasmussen (Eds.): DAGM 2004, LNCS 3175*, pp. 145–153, 2004.

[5] Mikolajczyk K. and Schmid C. (2004) – Scale & affine invariant interest point detectors. In *International Journal of Computer Vision* 60(1), pp. 63–86.

[6] Schober J.-P., Hermes T. and Herzog O. (2005) – PictureFinder: Description logics for semantic image retrieval. In *IEEE International Conference on Multimedia and Expo, 2005. ICME 2005*, pp. 1571–1574, Amsterdam, IEEE, July 2005.

[7] Porter M. F. (1980) – An algorithm for suffix stripping. In *Program*, 14(3), pp. 130–137.

[8] Schmid H. (1994) – Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

[9] Schmid H. (1995) – Improvements in part-of-speech tagging with an application to German. In *Feldweg and Hinrichs (Eds.): Lexikon und Text*, pp. 47–50.

[10] Frank E., Hall M. A., Holmes G., Kirkby R, Pfahringer B., Witten I. H. and Trigg L. (2005) – Weka - a machine learning workbench for data mining. In *Maimon and Rokach (Eds.): The Data Mining and Knowledge Discovery Handbook*, pp. 1305–1314, Springer.

[11] Hermes Th., Miene A. and Moehrke O. (2000): Automatic Texture Classification by Visual Properties. In *R. Decker and W. Gaul (Eds.): Classification and Information Processing at the Turn of the Millenium, Proc. 23rd Annual Conference Gesellschaft für Klassifikation e.V. 1999*, pp. 219–226, 10-12 March, Bielefeld, Germany.