AUTOMATED SUPPORT FOR A COLLABORATIVE SYSTEM TO ORGANIZE A COLLECTION USING FACETS

Kurt Maly[1]; Harris Wu[2]; Mohammad Zubair[1]; Victor Antonov[1].

[1] Department of Computer Science, Old Dominion University
Norfolk, VA, USA
e-mail: maly@cs.odu.edu; zubair@cs.odu.edu
[2] Department Information Technology and Decision Sciences, Old Dominion University, Norfolk, VA, USA
e-mail: hwu@odu.edu

Abstract

We are developing a system that improves access to a large, growing image collection by supporting users to collaboratively build a faceted (multi-perspective) classification schema. For collections that grow in both volume and variety, a major challenge is to evolve the facet schema, and to reclassify existing objects into the modified facet schema. Centrally managed classification systems often find it difficult to adapt to evolving collections. The proposed system allows: (a) users to collaboratively build and maintain a faceted classification, (b) to systematically enrich the user-created facet schema, and (c) to automatically classify documents into an evolving, user-managed facet schema. In this paper, we focus on (c), where we describe the approach to automatically classify documents into an evolving facet schema. We propose a learning-based system that periodically learns from manually classified images, and then classify new images accordingly.

Keywords: automated categorization; collaborative faceted classification; facets; support vector machines; image collections.

1. Introduction

Navigating a large growing collection of digital objects, particularly images and photographs, is challenging as keyword-based search has limited value. Typically this is addressed by classifying the collection using human experts in a centralized

way. Such a centralized approach often is prohibitively expensive and imposes a single-minded, static, rigid structure. However, staying away from static structure poses a challenge when navigating a large growing collection of digital objects, particularly images and photographs, where keyword-based search has limited value. Social tagging systems such as del.icio.us [1] and flickr.com [2] allow individuals to assign free-form keywords (tags) to any documents in a collection, and freely use each others' tags. While free, open and evolving, social tagging systems suffer from low quality and the lack of structure in tags. In absence of any guidance and structure, ambiguity and noise arise from the linguistic nature of tags such as polysemy, homonymy and synonymy.

We have proposed a system that improves access to a large, growing collection by supporting users to collaboratively build a faceted (multi-perspective) classification schema [3]. A faceted classification allows assignment of multiple classifications to an object, supporting multiple user perspectives in search and exploration. For example, e-Commerce sites such as eBay use faceted classification to organize their item collections by product category, price, color, brand name, etc. For collections that grow in both volume and variety, a major challenge is to evolve the facet schema, and to reclassify existing objects into the modified facet schema. Centrally managed classification systems often find it difficult to adapt to evolving collections. It is hoped that through users' collective efforts the faceted classification schema will evolve along with the user interests and thus help them navigate through the collection quickly and intuitively. The proposed system allows: (a) users to collaboratively build and maintain a faceted classification, (b) to systematically enrich the user-created facet schema, and (c) to automatically classify documents into an evolving, user-managed facet schema. In this paper, we focus on (c), where we describe the approach we have taken to automatically classify documents into an evolving facet schema. It should be noted here that the main approach relies on collaborative classification of images; however, for initial classification when we bring in new images into the collection we use the automated approach for initial classification.

In this paper, we focus on describing an approach to automatically classify documents into an evolving, user-managed facet schema. In our context, documents are images together with metadata such as description, title, and photographer. Even with a perfect facet schema, it would be useless unless most documents in the collection are classified into the schema and therefore accessible from browsing the faceted classification. Ideally documents should be "fully" and "correctly" classified, i.e. properly classified into all pertinent categories in all facets. This way browsing a faceted classification will have a high recall and precision of documents. For a large, fast growing collection, human efforts alone will not be sufficient to keep the documents "fully" classified.

The proposed approach creates a SVM (support vector machine)-based classifier for each category in the faceted classification schema. Users' manual classifications are utilized as training input, and all existing metadata for a given image are used as the feature set. As users continue to classify documents or affirm system-generated classifications, the classifiers are regenerated periodically using enlarged training sets.

The rest of the paper is organized as follows. In Section 2, we discuss past and related work. Section 3 discusses SVM based approach for classifying images with little metadata. In Section 4, we discuss implementation details. Finally, in the last section we have conclusions.

2. Past and Related Work

Categories represent a way content is organized into a structure that is both meaningful and traversable and which allows for objects to be easily retrieved for later usage. Images, in particular, need such organization because an image itself is not easily searchable for any specific information that is useful to the requestor. A commonly used approach is "tagging" images with keywords which can later be searched for. However, tags do not fully allow for browsing a collection by selecting and narrowing down collective criteria. It is categories that allow for multiple images that share common traits to be arranged together and, consequently, found together. Faceted categorization is an extension to the common category structure. Facets allow for an image to belong to more than one collective criterion (the facet). Within each facet, a regular, multi-tier category structure is developed. By allowing an image to possess several descriptive categorizations, browsing for specific needs becomes much easier. In addition, faceted categorization will ideally use far less categorization descriptors than a linear list of categories.

Traditionally, tagging and categorization in image classification systems have been the tasks of two dissimilar human groups. Tagging an image with keywords is generally the task of the users of the system. It represents their ability to associate what they are seeing with an idea or an object which they can easily recall later and search for. Very little input is needed by an administrative entity to collect and support such metadata. Faceted categorization systems, on the other hand, are typically created and maintained by a central entity. Facets and categories are created by the administrator or a group of experts and, with the exception of occasional changes, they remain very much static. As a result, many users' ideas of new classifications are not included in the scheme which can potentially reduce the intuitiveness of browsing the collection.

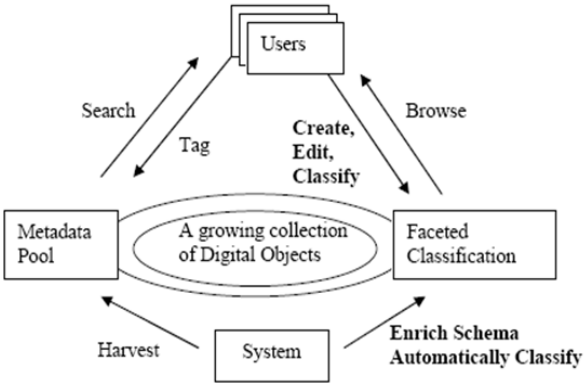


Figure 1. A high-level overview of the system.

Faceted classification consists of two components: the facet schema containing facets and categories, and the association between each document and the categories in the facet schema. As mentioned earlier, in order to create, maintain and enrich these two components, the proposed system (Figure 1) allows: (a) users to collaboratively build and maintain a faceted classification, (b) to systematically enrich the user-created facet schema, and (c) to automatically classify documents into an evolving, user-managed facet schema [4]. We have developed a Joomla-based browsing interface and a javascript-based classification interface [5] which allow users to create and edit facets and categories and view their contents by selecting only the desired categories in a fashion similar to eBay (Figure 2). They can also classify (or re-classify) images into faceted categories by intuitive point-and-click or dragging actions. Both interfaces, along with the server-side automated processes, are available for a regular installation of the Joomla content management system as they are intended as open source modules for the Joomla community. The entire system runs on top of a MySQL database which contains all categories, image metadata, as well as image-category associations.

Automated document classification has been an active field in the past decade. The basic non-learning methods, such as to categorize documents based on word matching between documents (using content or metadata) and category names, are quite limited for a faceted classification. For example, a document tagged with "apple" will match both the apple category as a fruit, and Apple computers. A variety of statistical learning methods have performed better than non-learning methods for documents classification.

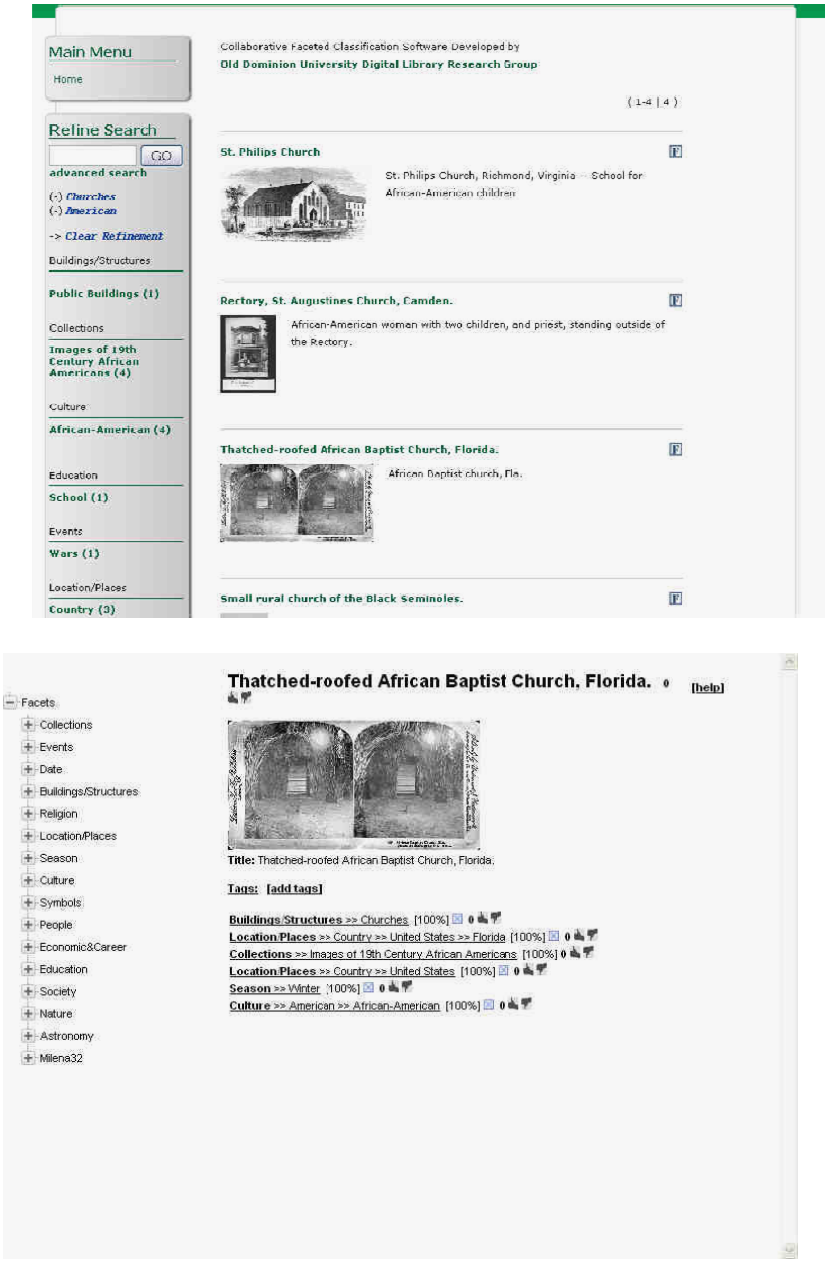


Figure 2. Browsing (above) and classification (below) interfaces.

These include regression models [6], nearest neighbor classifiers [7], decision trees [8], Bayesian probabilistic classifiers, inductive rule learning algorithms [9], neural networks [10], online learning approaches, support vector machines (SVM) [11],

genetic programming techniques, and many hybrid methods. There have also been a number of comparative studies of different classification techniques. One unique classification challenge to the proposed system, however, is that of the user-managed facet schema as a moving target. New classifiers need to be frequently added and the system re-trained. Incremental learning is desired. The classification technique also needs to fit the unique characteristics of a given document collection and available metadata input. Besides the choice of classification techniques, there is a timing issue for automated classification. As the facets evolve, many categories, especially the newly added ones, will change frequently. The system should wait until a category is "stabilized" before attempting to classify documents into the category. Also practically, learning-based techniques need a sizable training set to train the classifiers.

For this project SVM we chose and used the core functionalities of LibSVM, an integrated software for support vector classification, as backend algorithms to the system described above. LibSVM also supports regression and distribution estimation. The learning algorithm employed in LibSVM is a variation of the traditional SVM algorithm and has been described in [12]. The LibSVM package also includes cross-validation for model selection and probability estimates which we used in this project.

3. SVM Classification of Images with Little Metadata

3.1. Building SVM Classifiers

Since the Faceted Classification System is developed for collections of images, textual data for support vector machines is limited. Specifically, we have chosen collections of African-American historical images where each item is stored in a database with three metadata fields: title, description, artist/photographer and keywords. These three fields are to provide the input for training an SVM classifier. Since only the title field is required in the source collections when adding an item, the metadata can be rather poor in some cases. An example is shown in Figure 3.

In the image in Figure 3, the old photo has been added to the collection with a title, a description, and several keywords. It is clear that the photo represents a church with a thatched roof, that it is in Florida and it is a Baptist church. A text-based automated classifier such as SVM, however, tends to work with large data vectors and while the above information might be sufficient for a human visitor, the machine learning process will need additional data in order to be correctly ap-

plied later. Despite the existence of data in all three information fields, the metadata is repeated in each field which reduces the overall information. In other cases a portrait might be added only with the name of the person depicted and, if we are lucky, the name of the photographer or artist. A mere name is not enough for a classifier to assign categories to such an image.

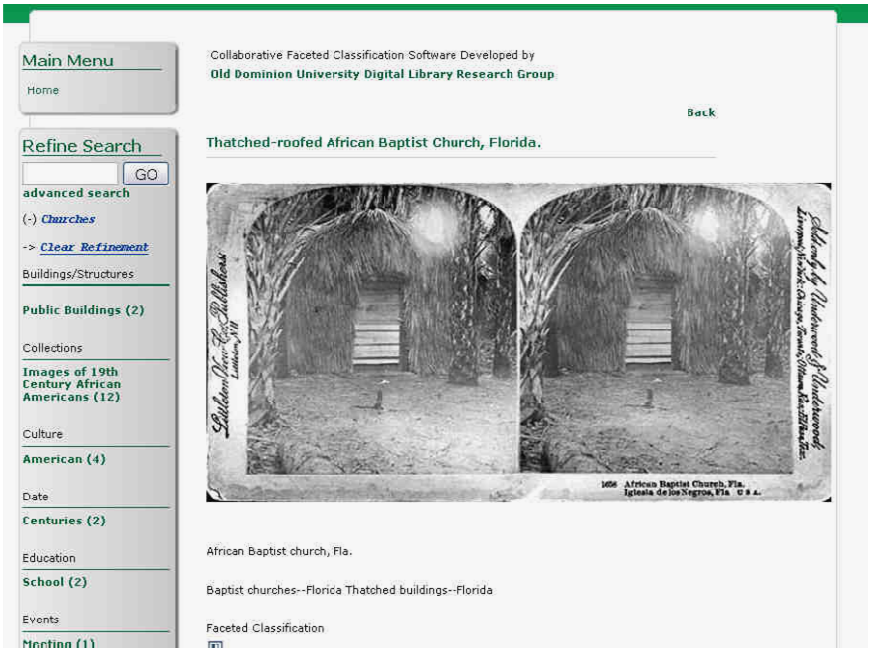
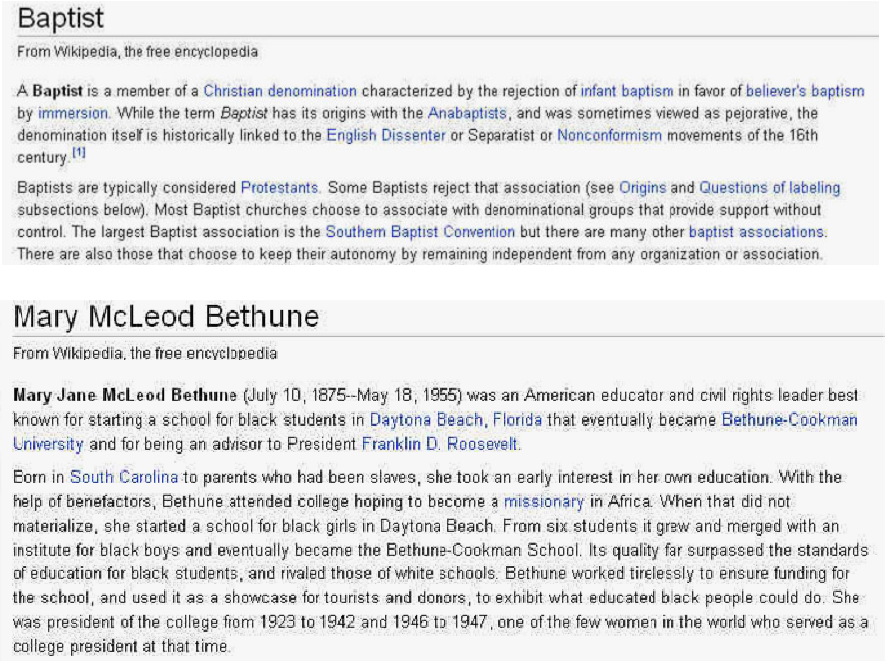


Figure 3. Example image with little metadata.

In order to test whether we can improve the training accuracy of SVM classifiers, we decided to expand this metadata set by harvesting information available on Wikipedia for the keywords available in these basic metadata fields. Classifiers trained in this manner will then be compared to a similar set of classifiers which were trained on the existing metadata alone. Since Wikipedia contains a vast body of knowledge, much unrelated information can be acquired through simple searches of common keywords. Since SVM eliminates common words, "stop words", we limited the Wikipedia search to proper nouns (names of people, places, etc.) found in the basic metadata fields of items in the database.

For the above example of the church image, the likely candidates for Wikipedia search terms would be "African", "Baptist", and "Florida". The latter two terms are specific, while the first one is broad. Through harvesting Wikipedia as a general knowledge source, we learn (or, rather, the SVM learns) that Florida is in the United States, that Baptist is a denomination of Christianity and that African can

relate to African or African-American culture. Furthermore, through the cross-section of this information and the existence of a dated image with similar metadata, the automated process might deduce a date for the image above. In the cases where an image only has a person's name, the latter's biography on Wikipedia will supplement the SVM with information on places and years, among other data. Figure 4 below gives an example of the sort of information which would be obtained from Wikipedia, both in a more general case and in the case when the search term is a personal name. In the latter case, the biography is beneficial to classification as the image of the collection only came with a description of what we see rather than who is the person we see. As indicated before, SVM removes stop words (which list can be expanded by the application writer to include common, irrelevant to the classification, words such as "Wikipedia" and "encyclopedia") from this result form Wikipedia.



Baptist

From Wikipedia, the free encyclopedia

A **Baptist** is a member of a Christian denomination characterized by the rejection of infant baptism in favor of believer's baptism by immersion. While the term *Baptist* has its origins with the Anabaptists, and was sometimes viewed as pejorative, the denomination itself is historically linked to the English Dissenter or Separatist or Nonconformism movements of the 16th century.[1]

Baptists are typically considered Protestants. Some Baptists reject that association (see Origins and Questions of labeling subsections below). Most Baptist churches choose to associate with denominational groups that provide support without control. The largest Baptist association is the Southern Baptist Convention but there are many other baptist associations. There are also those that choose to keep their autonomy by remaining independent from any organization or association.

Mary McLeod Bethune

From Wikipedia, the free encyclopedia

Mary Jane McLeod Bethune (July 10, 1875--May 18, 1955) was an American educator and civil rights leader best known for starting a school for black students in Daytona Beach, Florida that eventually became Bethune-Cookman University and for being an advisor to President Franklin D. Roosevelt.

Born in South Carolina to parents who had been slaves, she took an early interest in her own education. With the help of benefactors, Bethune attended college hoping to become a missionary in Africa. When that did not materialize, she started a school for black girls in Daytona Beach. From six students it grew and merged with an institute for black boys and eventually became the Bethune-Cookman School. Its quality far surpassed the standards of education for black students, and rivaled those of white schools. Bethune worked tirelessly to ensure funding for the school, and used it as a showcase for tourists and donors, to exhibit what educated black people could do. She was president of the college from 1923 to 1942 and 1946 to 1947, one of the few women in the world who served as a college president at that time.

Figure 4. Wikipedia content to be downloaded.

In order to build a SVM classifier, a minimum set of training data is required. This minimum depends on the application. As a result, a setting in the configuration file for the automatic classification system was introduced to choose only these categories which have at least a certain amount of item associations (item threshold). Furthermore, another setting was introduced which will control the number

of items used for the training process (positive training set). More often than not, the two settings will be equal but they can always be changed for testing purposes. In addition to the set of items from a given category, a set of equal (or greater when available) size is chosen among items which are associated with any other category but this one (negative training set). Items in both training sets are randomly chosen using a randomizing function.

As a next step in the process, documents (SVM expects a text document as input) for each item chosen for training had to be created. As we wanted to compare and maximize the effectiveness of using Wikipedia, we create two text documents for each image: one document containing basic metadata fields and the other includes information harvested from Wikipedia. In order to extract the proper nouns from the metadata fields, a stochastic parts-of-speech tagger is used. Relevant results are passed as search strings to Wikipedia where the article for the first search result returned (by default, Wikipedia returns the most relevant article as the first search result) is downloaded (Wikipedia-related metadata and image captions are stripped; raw text is downloaded only). Articles collected for a single item are collected into a single text document. Secondly, we create text documents for the same set of items but containing only metadata from the database. Since there are items that can participate in multiple training sets, both positive and negative, documents are created in common pools and are used when needed in training. Consequently, the need to create multiple copies of the same document is avoided, as well as the multiple calls to the same article in Wikipedia.

Early on during the development of this automatic classification system, however, we realized that content retrieved from Wikipedia might be extraneous if the item (image) in the database was aptly supported by metadata. Hence, SVM results based on Wikipedia texts were only to be used if the existing metadata was scarce. We introduced a threshold on the number of significant terms in the metadata. If this number was below the threshold level, both metadata- and Wikipedia-based SVM scores were used in the prediction by introducing a weighting function which gives more importance to keywords in metadata fields than the information harvested from Wikipedia. The weighing value was produced by the ratio of existing relevant terms in metadata and the term threshold. If metadata surpassed the threshold value or Wikipedia did not produce any significant terms, the weight was set to 1 for metadata-based SVM classifier and 0 for the Wikipedia based one (the latter value would not be used at all in prediction).

The first step of creating a SVM classifier is the creation of a collection model with all significant terms. Then this model is used along with the positive and negative training sets to create a statistical representation of the collection category.

3.2. Testing of SVM Classifiers

For the purposes of evaluating the SVM-based classification, we used a subset of five categories and all items within these categories. Human experts fully classified the documents (images) in this subset. Roughly half of the subset was used in training the classifiers (165 items). The rest (168 items) were used to test the effectiveness of the automated classification. Table 1 records the number of items used in training the SVM classifiers for the five categories:

Category Name	Positive Documents	Negative Documents	Average Wikipedia term count	Average metadata term count
Men	57	121	175	12
Women	53	110	202	13
Fashion	39	121	144	12
Famous Men	44	129	160	12
African-American Culture	27	131	219	12

Table 1: SVM training sets.

Once all the relevant collection models and classifiers have been created based on the training sets, the test set of 168 items was given to each of the five classifiers. When presented with an item (in the form of a text document), each classifier produced as output a number between 0 and 1, representing the probability of the item being associated with the respective category as a prediction value. To test our hypothesis on Wikipedia enhanced metadata we produced two SVM classifier for each category, one based on the metadata only and one based on metadata plus Wikipedia obtained terms. In order to evaluate the test results, we used standard measures. The first of these is precision, also known as positive predictive value. In terms of our test, this is the proportion of human-classified items which were correctly classified by the SVM-classifiers. The second measure is recall (sensitivity) which is defined as the proportion of actual positives which are correctly identified as such. The third is the F-measure which is the weighted harmonic mean of the two. Figure 5 shows the three formulae [13]:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}.$$

Figure 5. Information retrieval performance measures.

The above measures were computed for different values of two thresholds each. The first one is the term threshold that describes the minimum number of metadata terms after which no Wikipedia score will be used in the prediction. We selected five values: 10, 20, 30, 40, and 50 for the experiment. A term threshold of 0 indicates a document consisting of original metadata only. The second threshold is the SVM threshold on the prediction itself. In other words, this is the minimum value for a classification to be positive. This value can range from 0 to 1 and measures were taken at each decimal value (0.1, 0.2, etc.). The graph in Figure 6 shows the combined F-measure for all 5 SVM classifiers.

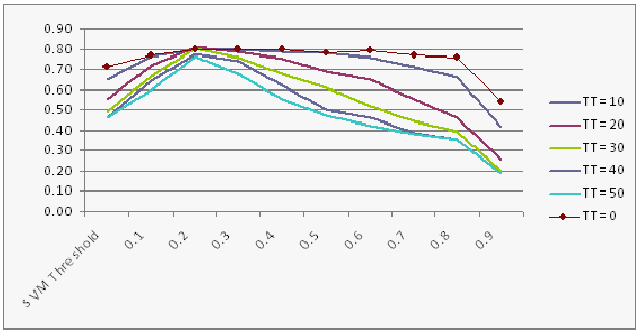


Figure 6. F-measure comparison.

The first observation made was that the F-measure peaks at SVM threshold of 0.3 in all five cases. While this comes as a surprising result, there is a reason for this value. The way the F-measure combines precision and recall, it does not take into account true negative machine predictions (cases where the SVM classifier did not place an item into a category the same way a human classifier would). Therefore, its value is significantly raised for large values of precision. At lower SVM threshold values precision is high, making the F-measure high as well. In order to capture the full extent of the SVM effectiveness, we introduced one more measure: the

specificity, or fall-out rate. It returns the ratio of true negative and all negative predictions. The specificity value reaches a ceiling value of 1 for SVM threshold values of 0.88 and above. However, at the same time recall is low and therefore F-measure values become low. At the peak of the F-measure (SVM threshold of 0.3), the specificity is 0.89, making this SVM threshold an appropriate value for document classification.

One cannot expect automated classification to be perfect and our faceted classification system will use this backend SVM process only as a supplement to user-generated classification. We consider the latter to be of expert quality (or close to). Therefore, users have the ability through the classification interface (Figure 2), to reassign item-category associations or to mark such associations for deletion as they see fit. In case users are unsure of the relevancy of a classification, be they human- or machine-generated, they are provided with a method to vote it down. Such uncertain classifications will be automatically removed by the system if negative voting is persistent. Conversely, associations that have been repeatedly voted on positively will stay. With the possibility of removing or changing a classification and voting on it, wrong classifications by the SVM automated process can be tolerated as such cases will gradually disappear in the collaborative environment.

The second observation from Figure 6 is that retrieval of additional terms from Wikipedia does not help document classification which came as a complete surprise to us. In fact, it is shown that it dilutes the already existing metadata with additional information which is probably not relevant for the classification purposes. In Figure 6, the higher the term threshold (TT), the more often Wikipedia text is used in the classification but the lower the F-measure is. In fact, for a term threshold of 0, the F-measure values are almost consistently high. The conclusion that can be drawn from this outcome is that an SVM classifier has a good overall performance even when existing metadata is not abundant and that while Wikipedia is a good source of general information and knowledge, it is not useful when it comes to classification improvement. However, it does not mean that the general idea of metadata augmentation is not a good one, it means we may have to rely on a more careful pruning of the result set and or have searches based on Boolean queries that involve also 'and' and not just 'or'.

We have considered using Google searches as an alternative to Wikipedia references. The Google search engine has its pros and cons over encyclopedic engines. First of all, Google indexes a much vaster set of web pages and the chance for positive results is therefore much higher. Google's engine allows for very specific results, especially when based on search strings of several relevant terms. For example, when using {"Indians", "tribute", "French", "Florida"} as the set of search terms, the top three results were linking to other occurrences of the same image in the web. However, there are several negative aspects to a Google-based search ap-

proach to augment the metadata. In the last example, where links to the same image were returned by the search engine, it is doubtful whether more information can be acquired for classifying purposes as the image has likely been added there with the same metadata as in our collection. In other experimental searches, the top results of Google were highly relevant to the search terms, albeit only textually as they were merely mentioned on the web page in question. As a result, if information was to be retrieved from there, it will be with low relevance towards describing the image. In spite of these uncertain results, in the future we will further investigate harvesting metadata from Google.

4. Implementation Details

In the process of developing the SVM-based classification method described in the previous section, we gathered statistics on performance and system requirements. In one of our early tests, the item threshold per category (least amount of items for a category to be "eligible" for a classifier) was assumed to be 20. This resulted in 13 categories being chosen for training. Using these parameters, the following observations were made regarding the entire training process, including Wiki and non-Wiki training:

	Wiki	**Non-Wiki**	**Total**
Average document collection creation time (20 pos + 20 neg) (min)	2:36	0:20	2:56
Average training time (min)	0:24	0:06	0:30
Document files used	520	520	1040
Actual files on system	449	462	911
Total size of documents (KB)	47,332	1860	49192
Size of SVM model files + logs (KB)	1776	428	2204

Table 3. SVM-based classifiers' performance statistics.

Using a common file pool for documents had a positive effect on space consumption as only 87.5% of the (13*40=) 520 documents were retained on the system. If each category had its own directories with documents, the remaining 12.5% would have been duplicates. In an earlier test run, when selecting items for the training sets was not random, only 249 documents (in both cases) were downloaded in this pool. This is due to the fact that many items would appear in more than one training set, especially the negative sets. Using this pooling strategy has had an effect

on the time performance as well, since documents were used by more than one category, but they were created only once. However, due to peculiarities of the software which prepares the special data files for SVM training (extraction of common terms and collecting them in vectors), temporary copies of the files are created in directories for each respective training set, which calls for a certain measure of temporary disk space to be available at all times when training is performed.

From the above observations, the following conclusions can be drawn on time performance and space requirements. The bulk of time in SVM training was spent in creating the document files. Extrapolating from results in the table above, creating a Wikipedia-enhanced document took an average of 4.5 seconds. Similarly, creating a metadata-only document took only an average of 0.675 seconds. If Wikipedia-enhanced training is to be used alongside metadata-only training, the space requirement will be quite significant. Using the data above, if each item from the collection had two files in the pool, one will need over 100 KB per item. For a collection of 100,000 items, this amounts to 10 GB of space. At the same time, if metadata-only training is used by itself, the space requirement for the same size of the database would only be around 400 MB. Similar conclusions can be made about the space requirements for the training and logging files created by the SVM process. Testing shows about 160 KB per category, or over 15 MB for 100 categories being trained. By extrapolation we can conclude that the final product (including both training methods) will require free space of 15 GB for a collection of 100,000 items and 100 categories (this includes the Joomla! software, MySQL database, drivers, and SVM-related files but does not include media files).

The Faceted Classification System is meant to be a collaborative effort and thus item associations and metadata can change rapidly. In order to accommodate this to the automatic classification, SVM classifiers need to be periodically retrained at intervals, set by the administrator. A category will have its respective SVM retrained if a certain number of items have been associated with it during this time interval (retrain threshold). Carefully setting these time periods will guarantee SVM classification based on recent data. Additionally, a new document will be created for an item if the existing document is older than a preset period of time. This will significantly decrease the time that the classification process spends in document creation which can otherwise be significant, as shown above.

The system that we are developing, including both its user interface and backend processes is in its post-development stage, there are still future steps and improvements to be considered. One such consideration is to follow evolution of the LibSVM suite. As of now this software has been used only in binary classification, that is, to determine whether an item belongs to a given category or not. Hence, each category in the schema needed its own classifier. However, LibSVM is said to

support multi-label classification or the possibility to train a classifier for a set of categories and to predict the probability of an item belonging to each category in this set. This has the potential of significantly reducing the length of the training period while producing similarly good results. So far support for such classification is not native to LibSVM and third-party software is needed to break down the process into binary classification. In addition, we shall consider and test a form of retraining threshold which is percentage-based rather than count-based which will be "fairer" to slowly growing categories. Thirdly, we are exploring ways to establish feedback loops so that parameters of SVM can be adjusted, based on information obtained from changes humans have made to classifications originally made by SVM.

5. Conclusion and Discussions

A collaborative user effort concentrated on a combination of keyword assignment and faceted categorization has the potential to greatly improve image classification to support search/browse and subsequent retrieval. However, a downturn of a faceted retrieval system is that all items in the collection need to be categorized in order to be found quickly. Users of the system might not have the knowledge or time to place images into appropriate categories and/or create new categories or change the existing ones. Therefore, save for expert administrative effort, an automated classification method is required which will supplement the user effort and will provide the basis for the evolution of the overall classification.

In the past support vector machines have proved to be very good methods for text classification. In the case of visual media such as images, classification of this type can only work with existing metadata for the image item. In this paper we have presented a SVM-based approach to the problem. A significant portion of the image collection for our test bed had only minimal metadata and we proposed to solve this scarcity problem by augmenting the metadata. Our thesis was that using the proper nouns of the metadata as search keys for Wikipedia and using the text of the first result was an effective way to improve the SVM classification. To our surprise the thesis was proven wrong and we ended up using the augmentation only when metadata consisted of only a single piece. We are exploring other supplementary metadata augmentation options such as Google search results and semantic analysis of existing metadata terms in combination with Google or Wikipedia searches.

We have developed an automated backend system that will classify all unclassified items of the original collection as long as there are sufficient items in a cate-

gory that have been already manually classified. Transparently to the user, our system will monitor all categories and periodically reclassify and or retrain. The basis for the time when this happens is the number of user added items to the collection and user classifications of items.

Besides metadata augmentation, feedback loops for user and SVM classification, choices for various thresholds, we shall also explore a KNN (k-nearest neighbors) approach to the automated classification problem. One great advantage of KNN would be that it requires no training process. The KNN algorithm calculates distances between items based on their term vectors. [7] has explored the method in depth and [14] have shown the effectiveness of the KNN algorithm in image classification in comparison to learning based algorithms like SVM. However, the latter concentrates on using image descriptors as data vectors. We shall explore the use of KNN when using textual metadata.

References

[1] GOLDER, S.A. and HUBERMAN, B.A. The Structure of Collaborative Tagging Systems. Information Dynamics Lab, HP Labs, 2005.

[2] MARLOW, C. et al. Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, May 2006. [Online]. Available at www.danah.org/papers/WWW2006.pdf (March 2009).

[3] MALY, K., WU, H. and ZUBAIR, M. Harvesting social knowledge from folksonomies. *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*. New York: ACM, 2006, p. 111-114.

[4] MALY, K., WU, H. and ZUBAIR, M. Collaborative classification of growing collections with evolving facets. *HT '07: Proceedings of the Eighteenth Conference on Hypertext and Hypermedia*. New York: ACM, 2007, p. 167-170.

[5] More information on the Faceted Classification System and module downloads can be found at http://facet.cs.odu.edu/.

[6] YANG, Y. and CHUTE, C.G. A linear least squares fit mapping method for information retrieval from natural language texts. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING 92) - Volume 2*. Morristown, NJ: Association for Computational Linguistics, 1992, p. 447-453.

[7] CREECY, R.H. et al. Trading mips and memory for knowledge engineering: classifying census returns on the connection machine. In *Communications of the ACM - Volume 35, Issue 8*. New York: ACM, 1992, p. 48-63.

[8] MOULINIER, I. Is learning bias an issue on the text categorization problem? In *Techincal report, LAFORIA-LIP6*. Paris: Universite Paris, 1997.

[9] COHEN, W.W. and SINGER, Y. Context-sensitive learning methods for text

categorization. In *SIGIR'96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Zurich: ACM, 1996, p. 307-315.

[10] WIENER, E. et al. A neural network approach to topic spotting. In *Proceedings of SDAIR'95, the Fourth Annual Symposium on Document Analysis and Information Retrieval*. Las Vegas: 1995, p. 317-332.

[11] JOACHIMS, T. Text categorization with support vector machines. In *Proceedings of 10th European Conference on Machine Learning*. London: Springer-Verlag, 1998, p. 137-142.

[12] FAN, R.-E. Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research – Volume 6*. December 2005. [Online]. Available at: http://www.jmlr.org/papers/volume6/fan05a/fan05a.pdf .

[13] Information Retrieval. In *Wikipedia, the free encyclopedia*. March 2009. [Online]. Available at: http://en.wikipedia.org/wiki/Information_retrieval (March 2009).

[14] BOIMAN, O. et al. In Defense of Nearest-Neighbor Image Classification. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Anchorage: IEEE, 2008, p. 1-8.

June 2009
Printed on demand
by *"Nuova Cultura"*
www.nuovacultura.it

Book orders: ordini@nuovacultura.it