

Classifier Hypothesis Generation Using Visual Analysis Methods

Christin Seifert, Vedran Sabol, and Michael Granitzer

Know-Center Graz, Austria
{cseifert,vsabol,mgrani}@know-center.at
<http://www.know-center.at>

Abstract. Classifiers can be used to automatically dispatch the abundance of newly created documents to recipients interested in particular topics. Identification of adequate training examples is essential for classification performance, but it may prove to be a challenging task in large document repositories. We propose a classifier hypothesis generation method relying on automated analysis and information visualisation. In our approach visualisations are used to explore the document sets and to inspect the results of machine learning methods, allowing the user to assess the classifier performance and adapt the classifier by gradually refining the training set.

Keywords: Text Categorisation, Visual Analysis.

1 Introduction

In today's information-driven world new documents, such as news, scientific publications, technical reports or patents are produced at an astonishing rate. Frequently the need arises to supply recipients with particular topics of interest with relevant new documents. This task can be automated by using a classifier trained to recognise documents which are relevant to a particular topic, i.e. classify the documents to the corresponding topical category. Obviously, accurate automatic approaches for large data sets are highly desirable. As the performance of a classification model strongly depends on the training data, the classifier needs training data that is representative for the data set. All categories must be sufficiently covered with examples which preferably contain no contradictions. The situation is further aggravated in dynamic data sets where the problem of keeping the training set up to date may arise as new topics appear, vocabularies drift, or interests of the recipients gradually change. In this paper we present a work in progress which attempts to address these issues by a visual analytics-based approach, where automated analysis is combined with information visualisation to unite the strengths of high-speed computer processing with immense pattern recognition capabilities of the human visual apparatus.

We propose a classifier hypothesis generation method combining unsupervised and supervised machine learning methods complemented by human involvement via visual analysis GUI components. Users' general knowledge and intuition are

decisive factors in steering the training set definition process and providing feedback to the system. Definition of the training set for a category usually begins by selecting candidate documents using a keyword search. In our approach the potentially large and diverse search result set is analysed by a clustering algorithm and presented to the user in an Information Landscape visualisation. The insights the user gains while exploring the topical structures in the landscape are used for specifying the training set of the classifier. After the classifier was trained and new documents were classified, classification results with high confidence values may be gradually added to the training set if deemed good by the user. A classifier visualisation is used to (periodically) assess the training set of the classifier and identify classes and documents where the confidence values produced by the classifier do not appear satisfactory. These documents can be analysed again using clustering and Information Landscape with the goal of improving and tuning the training set assignments. Requirements for the proposed concept, which is currently in a prototype stage, are derived in cooperation with real world users who, within their productive environments, need to supply relevant documents to recipients interested in particular topical categories. Obviously, correct assignments of documents to categories as well as precise category definitions are central to the task.

The remainder of the paper is structured as follows: In section 2 we will briefly review related work, section 3 describes our approach and the required building blocks in detail. Section 4 is dedicated to a usage scenario on the Reuters-21578 text collection. We conclude and given an outlook to future work in section 5.

2 Related Work

The interdisciplinary research field of visual analytics focuses on reasoning facilitated by interactive visual interfaces [22]. It is a combination of automated discovery and interactive visualisation [9] used for understanding patterns in large data sets and discovering previously unknown facts and relationships. A core challenge in Visual Analytics is the analysis of massive repositories of unstructured texts [22]. The Information Landscape is a visual representation used to visualize complex relationships in large data sets. It has been successfully applied to convey topical relatedness in text document data sets in systems such as IN-SPIRE [11]. Information Landscapes have been extended to accommodate hierarchically organized text repositories, for example in InfoSky [3] and [17]. Mayer et al. [13] propose a map-based interface to large text collections based on self-organising maps (SOMs). This SOM-based visualisation also provides access to hierarchical topical clusters and could be used in our combined approach interchangeable with the Information Landscape.

Interactive machine learning offers a way to integrate background knowledge of domain experts into data mining models, in case the visualisations are designed appropriately [21]. For the task of classification, several visualisations have been proposed, most of them for specific classification models like for the Naive Bayes classifier [5]. Visualisations suitable for arbitrary classifiers are either restricted

to binary classification tasks, i.e. the visualisation based on self-organizing maps in [16], or applicable to only a small number of classes like the cobweb-based visualisation of the classifier decision quality [6]. A visualisation for arbitrary classifier has been proposed in [19], however the applicability to massive text data sets has not been shown yet. Interactive approaches include an interactive decision tree [12] and support vector machine construction [15]. However these approaches are not targeted at laypersons.

The Nora project [14] aims at constructing classifiers from large text repositories by letting the user label text documents with one of two category labels. The authors claim that their user interface design can be easily adapted to other text classification task where users can create training sets for classifiers. We think that an explicit representation of the overall classifier and its quality is missing in the interface. Another interactive text classification application has been proposed in the field of intrusion detection [4] where the authors combined a Naive Bayes classifier with a colour coded representation of text, and again let the user interactively label the incoming connection as benign or malicious. Here again, the user has no overview of the classifier and its quality, making it hard to assess the suitability of the classification model for the task.

Besides visual approaches, automatic approaches exist for creating classifiers from large data sets, namely semi-supervised learning techniques [24] and active-learning techniques [20]. Semi-supervised models need to be carefully designed and adapted to the problem structure in order to improve classifier performance compared to purely supervised learning [24]. Thus, semi-supervised approaches are not suitable for our application domain, because we can not make a-priori assumptions about the classification problem (e.g., estimated model complexity, data distribution). Active-learning techniques generate new training data by asking the user to label the data items for which the classifier is least confident [20]. Such techniques can be combined with our visual approach to increase the classifier's performance, once the categories are defined. For our application domain, the categories may evolve over time and thus a pure active learning approach is not applicable.

3 Combined Approach

Our approach to user-centred hypothesis generation for text categorisation is summarised in figure 1(a). Automatic techniques (depicted in dark grey boxes) alternate with users' analysis and actions (depicted by the light-grey boxes, tagged with the symbolic user). As figure 1(a) shows the approach is characterised by an analysis-action loop which is terminated when the user is satisfied with the classification hypothesis.

In detail, our approach consists roughly of the following steps:

1. **Search** (optional): Automatically finding potential documents of interest using keyword search.
2. **Pattern analysis in the Information Landscape**: Understand topical structures in the data set and select documents covering relevant topics. If

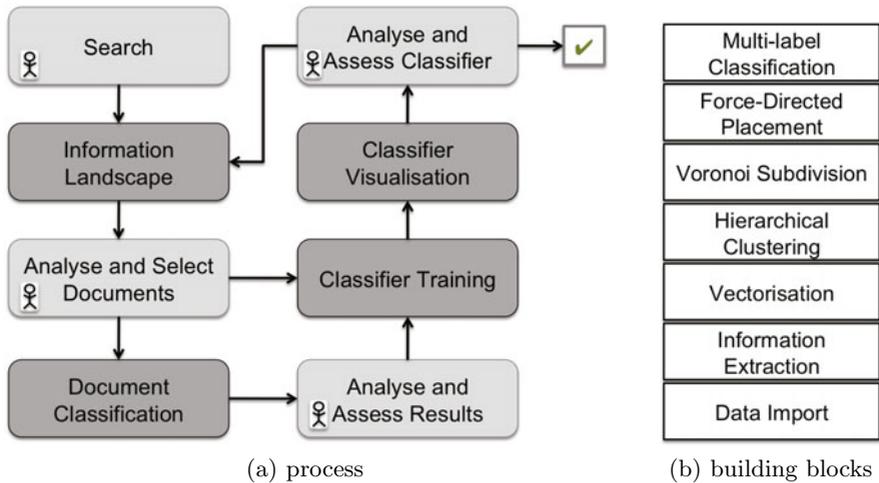


Fig. 1. Overview of the combined user-centred clustering and classification approach

no search was performed before the whole document set is chosen for the Information Landscape.

3. Building the classification hypothesis:

- **Training:** Use selected documents to train a new category or to modify (add/remove documents from) an existing one.
- **Classification:** Classify selected documents and, optionally, add documents with high confidence values to the training set of the assigned category.
- **Visualisation:** Inspect the classifier and, if deemed necessary, select documents for further analysis and refinement in the Information Landscape (step 2).

The building blocks necessary to implement the described approach are depicted in figure 1(b) and briefly described in detail in sections 3.1- 3.7. Note that the majority of the applied algorithms and visualisations are implemented within the KnowMiner knowledge discovery framework [10] and VisTools visualisation library [17].

Applying clustering and classification methods in the document space will result in a structure as depicted in figure 2. This structure contains two, in general orthogonal dimensions: the clustering tree and a forest of tree stumps imposed by the classification. The first structure devised by the clustering algorithm is visualised in the Information Landscape. Note, that the tree-like structure imposes a non-overlapping division of of the document set, i.e., a document only belongs to one of the next-to-bottom level clusters. The second structure, the decision tree stumps were imposed by the classification algorithm. The categories (on the right-hand side of figure 2) further divide the document set - independent of the cluster hierarchy. As we apply a multi-label classifier this second structure

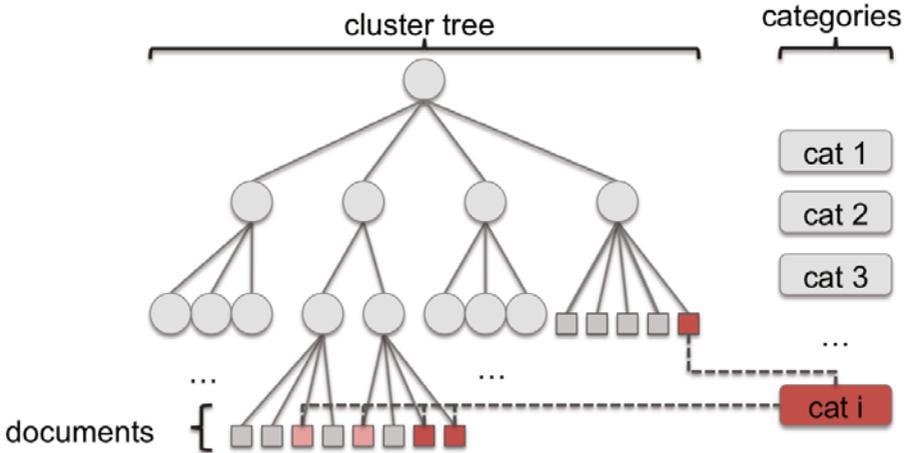


Fig. 2. Structure of the document space: hierarchical cluster tree superimposed by categories modelled by the classifier. Dark red documents are training documents for one class, light red are the documents classified by the classifier.

is in general not a partitioning of the document set, i.e., one document can belong to more than one category. The clustering hierarchy corresponds to topical structures implicitly present in the document space which are useful for gaining insight into the data, whereas the division imposed by the classifier corresponds to the structure explicitly defined by the application domain and users' needs.

3.1 Document Preprocessing and Indexing

Before any of the analytic algorithms and visualisations can be applied on text documents, these documents need to be preprocessed and transformed into a term space representation. Each document is represented by term vector where components of the vector are the frequencies (occurrence counts) of terms in the document. To extract relevant terms from a document we apply a part of speech tagger to identify nouns which are subsequently stemmed and stop word-filtered. As vectorisation of text documents may be quite time consuming, raw vectors are stored so they can be quickly retrieved when needed for processing by an algorithm. Also, all documents which are imported and stored in their vectorised form are also indexed so they can be quickly retrieved using full text search.

3.2 Hierarchical Clustering

Clustering is an unsupervised machine learning technique which partitions a given set of items, in our case text documents, into subsets (clusters) of related items. Documents assigned to the same cluster are similar to each other according to a similarity function. We apply the k-means algorithm recursively using the cosine similarity measure, which is known to perform well for text data [23].

Recursive application of k-means creates a hierarchy of clusters and sub-clusters where the leaves of the created cluster tree correspond to single documents. Each cluster in the hierarchy is labelled with the highest weight terms (i.e. extracted keywords) of the underlying documents. A mechanism for splitting and merging of clusters attempts to guess the "optimal" number of child-clusters, whereby the number of children is limited for usability reasons (i.e. to avoid scanning of long lists). The split-and-merge strategy also prevents the degeneration of the hierarchy. The resulting cluster hierarchy is suitable for browsing of the document set and we also refer to it as "virtual table of contents".

3.3 Projection Algorithm

The projection algorithm performs a dimensionality reduction of the high dimensional term space. In the resulting 2D visualisation space high-dimensional relationships (i.e. topical similarities) are preserved as well as possible so that topically similar documents (and clusters) are placed close to each other while dissimilar ones are positioned far apart. The algorithm [17] is a combination of hierarchical clustering techniques, force-directed placement and spatial tessellation, which proceeds recursively along the cluster hierarchy: First, the top level clusters are placed inside a rectangular area using a simple force-directed placement algorithm. The similarity of the centroids is calculated as the cosine similarity in the vector space representation of the centroids. After the top-level centroids are placed a Voronoi subdivision is calculated using the centroids as generator points for the Voronoi regions. The sub-clusters of a specific cluster are recursively projected inside the Voronoi region of a this cluster. The leaves (documents) of the hierarchy are placed within the Voronoi area of their parent cluster using the same force-directed placement method.

3.4 Multi-label Text Classification

The purpose of text classification algorithms is assigning category labels to previously unseen documents. Classification is a supervised machine learning technique, meaning that the algorithms learns the categories from a training set comprising of document-category pairs. Especially in text classification tasks single-label classification is not sufficient, meaning that each document may belong to more than one predefined category. We apply an adapted K-Nearest Neighbour (KNN) algorithm [2,1] for multi-label classification. As similarity we use the cosine similarity on the TF-IDF weighted vector-space representation of the documents. The output of the classifier for each classified document is a list of categories accompanied with a confidence value for each category. The visual analytics application is in principle independent of the specific classifier, as long as multi-label classification is supported. We use a KNN implementation to accommodate the dynamic nature of document repositories which are often growing at a fast rate. KNN training performance is suitable for frequently changing training sets where documents defining a training set of category are added (or removed) fairly often, and where categories need to be reorganised from time to time.

3.5 Information Landscape

The Information Landscape is a visual representation based on the geographic map metaphor. It is used to visualize complex relationships in large data sets by conveying relatedness through spatial proximity. We use an Information Landscape to visualise projection and clustering results provided by the algorithms described in sections 3.2 and 3.3. The cluster hierarchy is represented by nested polygonal areas which were generated by the Voronoi area subdivision. A region of the landscape corresponding to a cluster is labelled by highest weight terms of cluster's centroid. Documents, which are placed at the bottom of the hierarchy, are visualised as dots. Hills represent regions populated by a large number of topically related documents. They are separated by lower areas or see which represent sparsely populated regions.

The landscape (see figure 3) is an interactive GUI component designed for explorative navigation in the visualised data set. It adheres to the well-known information visualisation mantra (“overview first, zoom and filter, details-on-demand“) by providing an overview of the whole data set and, when required, offering insight into at finer levels of detail. Labels are useful both for orientation and navigation - clicking on the label will trigger a short animated “flight” to the corresponding cluster and reveal the areas and labels of its sub-clusters. At the finest level of details information on individual documents will be displayed. This provides an adaptive level of detail which is always adjusted to the zoom level and the area currently explored by the user. Free zooming, panning, rotating and tilting are also available. The landscape offers several mechanisms for selecting documents: using a lasso selection tool, single selection through mouse clicks, depending on the cluster membership, and using search. Selected documents are enlarged and/or displayed in a different colour.

3.6 Classifier Visualisation

The visualisation of the classifier provides an overview of the quality and the model of arbitrary classifiers. It is described in detail in [18] for single-label classification tasks. In our application data items may belong to more than one category (multi-label classification). For each classified document the classifier delivers category assignments consisting of pairs in the form *category label, confidence*, where confidence is a real number between 0 and 1 (highest). In the visualisation the categories are equally distributed on the circumference of a circle while the items are attracted to the categories according to their confidence values. This means that items placed near the categories belong to this category with high confidence. On the contrary, items placed in the centre of the circle are assigned to more than one category with high confidence. Thus, the visualisation gives an overview of the item distributions over categories. If most of the items are placed in the centre, this indicates a strong multi-label classification model. In the contrary, if most of the items are placed near the categories, it is an indication of a predominantly single-label classification model. As mentioned in [18] in the classifier visualisation the placement of the items is ambiguous.

This ambiguity is resolved by user interaction: Moving the mouse over a data item displays the corresponding assignments of the classifier, highlighting the category for which the classifier is most confident. Furthermore, on mouse over the content of the data item is shown allowing the user to assess the classifier’s decision. The classifier visualisation offers single selection on mouse click as well as a lasso selection to select several similarly classified items

3.7 Interaction Mechanisms

Table 1 shows all tasks that can be performed using both interfaces. Some tasks, like “delete items” can be performed in both user interfaces. Other tasks can only be performed in one of the interfaces. For instance, “delete class” is only possible from the classification window, since a class is not visible as such in the Information Landscape.

Table 1. Overview of the tasks that can be performed from the interfaces Information Landscape (IL) and classifier window (CW)

task	invoked in		results in
	IL	CW	
create category	✓		a new category is created from the selected documents and the classifier is retrained
delete from category	✓	✓	selected documents are deleted from the category and the classifier is retrained
delete category	✓ ¹	✓	the selected category is deleted and the classifier is retrained
classify		✓	the selected documents are classified and the output of the classifier is presented to the user
inspect documents	✓	✓	an Information Landscape is built from the selected documents
inspect category		✓	an Information Landscape is built from the training documents of the selected category
inspect classifier		✓	shows the classifier visualisation for the training data

4 Usage Scenario

We performed our experiments on the Reuters 21578 text collection. The hierarchical clustering results in 10 top-level clusters as shown in Figure 3. This structure of the information space is purely unsupervised. The Information Landscape in Figure 3 gives an overview of the entire text collection, showing clusters of similar documents and associated labels. The user can investigate the cluster hierarchy and get an insight in the overall content of the collection. The user

¹ Category deletion can not be explicitly invoked from the Information Landscape, but a category is automatically deleted if all its training documents are deleted.



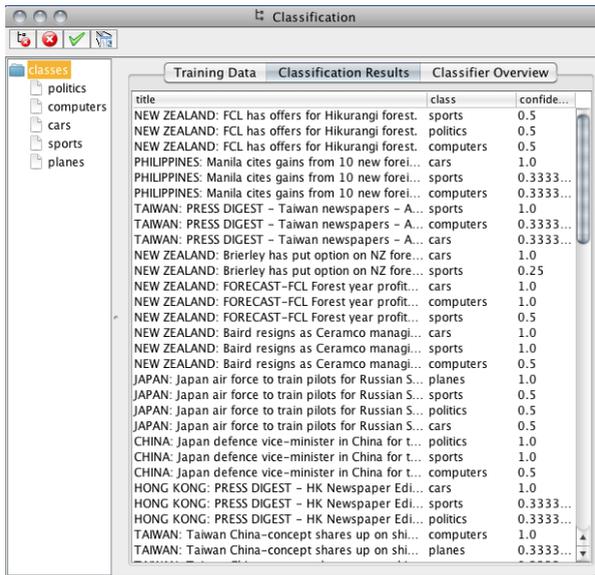
Fig. 3. Selection of documents in the Information Landscape

might be interested in other partitions of the data set which are not detected by the unsupervised methods. For example, the user might want to distinguish the categories “politics”, “computers”, “cars”, “sports” and “planes”. First, these categories are not explicitly represented, they only exist as a mental model in the user’s mind. While investigating the Information Landscape the user might come across documents that belong to one of these categories. The user can then select these documents (as shown in figure 3) and can create a new category from the selected documents. In the background, the selected documents are added to the classifier as new training data for the specific category (if the category does not exist in the classifier yet, it will be created). After repeating the steps “investigation” and “adding training documents to the classifier” the user might have found example documents for each of the categories of interest. He or she might then be interested of the current available documents for each category and the quality of the classifier that he or she has implicitly generated. This information is provided by the classification window. The training document for each class are presented as a list to the user. If the user detects wrongly assigned documents for a category he or she can simply remove them from the list and the classifier is retrained on the reduced training data set. For assessing the overall classifier quality the user can switch to the classification visualisation view as shown in Figure 5. In the figure, it can be seen that there are many documents belonging to more than one class (the central area). Only for the class “car” there are documents belonging to no other class. Further, there are some documents belonging to exactly 2 classes, these are the documents lying on the imaginary line between the “cars” and the “sports” rectangle as well as on the imaginary

line between “cars” and “computers”. The user might investigate the content of the interesting documents by moving the mouse over the items and eventually discover misclassified items. If user discovers that the classification model is not in line with his or her mental model of the categories, e.g., that the categories should be more distinct (i.e. lesser documents in the centre of the visualisation), the user could select the conspicuous documents and generate a new Information Landscape in order to further investigate them. Similarly a new Information Landscape can be generated for all training documents of one category. The resulting landscape is shown in figure 6 This might lead to further insights and actions, for instance finding and deletion wrongly assigned training documents. After cleaning up the classifier by consolidating the training sets, the user might be interested if there are more documents inside the collection that fit into these categories. Back in the Information Landscape he or she then selects documents and gets them classified. The classification result for the documents selected in figure 3 is shown in figure 4. Then the classification results can be investigated and, in the case the classifier correctly classified the documents, can be added to the trainings data set.

5 Discussion

Generating classifier hypothesis for large dynamic text data repositories is a challenging and time-consuming task. We described our work in progress, which combines automatic and visualisation-based approaches. The user is presented an interactive visualisation of the text collection, the Information Landscape, which



title	class	confide...
NEW ZEALAND: FCL has offers for Hikurangi forest.	sports	0.5
NEW ZEALAND: FCL has offers for Hikurangi forest.	politics	0.5
NEW ZEALAND: FCL has offers for Hikurangi forest.	computers	0.5
PHILIPPINES: Manila cites gains from 10 new forei...	cars	1.0
PHILIPPINES: Manila cites gains from 10 new forei...	sports	0.3333...
PHILIPPINES: Manila cites gains from 10 new forei...	computers	0.3333...
TAIWAN: PRESS DIGEST - Taiwan newspapers - A...	sports	1.0
TAIWAN: PRESS DIGEST - Taiwan newspapers - A...	computers	0.3333...
TAIWAN: PRESS DIGEST - Taiwan newspapers - A...	cars	0.3333...
NEW ZEALAND: Brierley has put option on NZ fore...	cars	1.0
NEW ZEALAND: Brierley has put option on NZ fore...	sports	0.25
NEW ZEALAND: FORECAST-FCL Forest year profit...	cars	1.0
NEW ZEALAND: FORECAST-FCL Forest year profit...	computers	1.0
NEW ZEALAND: FORECAST-FCL Forest year profit...	sports	0.5
NEW ZEALAND: Baird resigns as Ceramco managi...	cars	1.0
NEW ZEALAND: Baird resigns as Ceramco managi...	sports	1.0
NEW ZEALAND: Baird resigns as Ceramco managi...	computers	0.5
JAPAN: Japan air force to train pilots for Russian S...	planes	1.0
JAPAN: Japan air force to train pilots for Russian S...	sports	0.5
JAPAN: Japan air force to train pilots for Russian S...	politics	0.5
JAPAN: Japan air force to train pilots for Russian S...	cars	0.5
CHINA: Japan defence vice-minister in China for t...	politics	1.0
CHINA: Japan defence vice-minister in China for t...	sports	1.0
CHINA: Japan defence vice-minister in China for t...	computers	0.5
HONG KONG: PRESS DIGEST - HK Newspaper Edi...	cars	1.0
HONG KONG: PRESS DIGEST - HK Newspaper Edi...	sports	0.3333...
HONG KONG: PRESS DIGEST - HK Newspaper Edi...	politics	0.3333...
TAIWAN: Taiwan China-concept shares up on shi...	computers	1.0
TAIWAN: Taiwan China-concept shares up on shi...	planes	0.3333...

Fig. 4. Classification results for selected documents

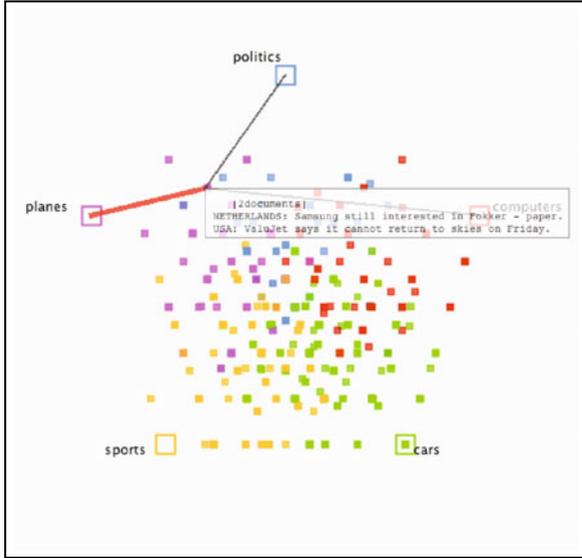


Fig. 5. Overview of the classifier’s training data set

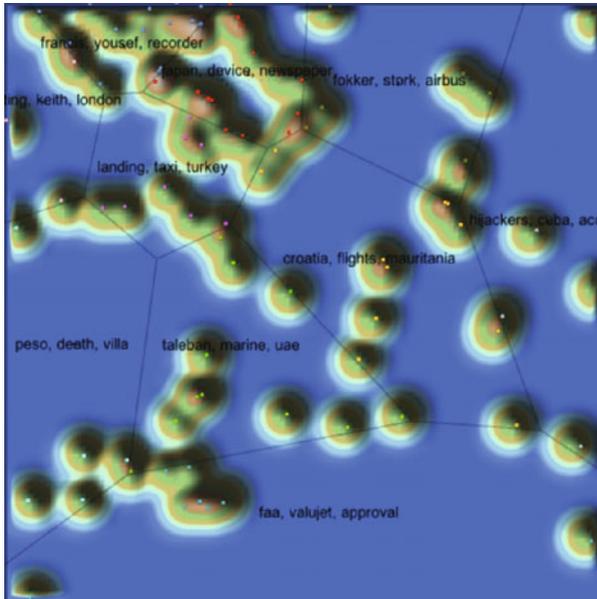


Fig. 6. Information landscape generated by selected documents (training documents for class “plane”)

is useful for gaining insights into topical structures present in the data set. The newly discovered information is useful for defining the training set of the classifier. The resulting classifier can be evaluated by the means of a classifier visualisation and refined further if necessary. We see clear advantages of our approach in the case when the categories are not pre-defined, but emerge during investigation of the document set. However, we also believe that the Information Landscape is useful for analysis and improvements of existing training sets, for example when the quality of the classifications is deemed unsatisfactory by the user.

6 Future Work

As our approach relies heavily on visualization and user interaction, usability evaluation of selected components would be necessary to discover eventual shortcomings. Some components, such as the Information Landscape, have already been evaluated in formal usability experiments [3,17]. Evaluation of the classifier visualisation component and its interaction with the Information Landscape appears as a natural next step. Even more importantly, an evaluation of the overall effectiveness of our approach is necessary to assess its practical applicability. As the development of our method is being driven by real world scenarios, we plan to test the effectiveness of our method with pilot users who in their daily work deal with assigning documents to topical categories. To obtain objective performance figures, we will compare the performance (expressed by quality and productivity indicators) of our classifier-based method to the solutions currently employed by the pilot users. These include either reading of documents and manually assigning them to categories, or when the amount of documents is prohibitively large, constructing complex Boolean search queries to narrow down the document set. Also, subjective user satisfaction should be evaluated through questionnaires and by collecting user remarks, which will help us identify main problems sources and provide hints on how to deliver remedies. Further direction for development of the system include primarily incorporating active learning methods to improve the classifier's performance once the categories are defined. Also we plan to integrate other classification models, such as Support Vector Machines [8] and the Class-Feature-Centroid classifier [7].

Acknowledgments. The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency (FFG).

References

1. Aha, D.W.: Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *Int. J. Man-Mach. Stud.* 36(2), 267–287 (1992)
2. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. *Mach. Learn.* 6(1), 37–66 (1991)

3. Andrews, K., Kienreich, W., Sabol, V., Becker, J., Droschl, G., Kappe, F., Granitzer, M., Auer, P., Tochtermann, K.: The InfoSky Visual Explorer: Exploiting hierarchical structure and document similarities. *Information Visualization* 1(3-4), 166–181 (2002)
4. Axelsson, S.: Combining a bayesian classifier with visualisation: Understanding the IDS. In: *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, pp. 99–108. ACM Press, New York (2004)
5. Becker, B.G.: Research report: Visualizing decision table classifiers. In: *Information Visualization*. IEEE Computer Society Press, Los Alamitos (1998)
6. Diri, B., Albayrak, S.: Visualization and analysis of classifiers performance in multi-class medical data. *Expert Systems with Applications* 34(1), 628–634 (2008)
7. Guan, H., Zhou, J., Guo, M.: A class-feature-centroid classifier for text categorization. In: *Proceedings of the 18th international conference on World Wide Web (WWW)*, pp. 201–210. ACM, New York (2009)
8. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) *ECML 1998*. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
9. Keim, D.A., Mansmann, F., Oelke, D., Ziegler, H.: Visual analytics: Combining automated discovery with interactive visualizations. In: Boulicaut, J.-F., Berthold, M.R., Horváth, T. (eds.) *DS 2008*. LNCS (LNAI), vol. 5255, pp. 2–14. Springer, Heidelberg (2008)
10. Klieber, W., Sabol, V., Muhr, M., Kern, R., Granitzer, M.: Knowledge Discovery using the Knowminer framewor. In: *Proceedings of the IADIS International Conference on Information Systems*, pp. 307–314 (2009)
11. Krishnan, M., Bohn, S., Cowley, W., Crow, V., Nieplocha, J.: Scalable visual analytics of massive textual datasets. In: *IEEE International on Parallel and Distributed Processing Symposium, IPDPS 2007*, pp. 1–10 (March 2007)
12. May, T., Kohlhammer, J.: Towards closing the analysis gap: Visual generation of decision supporting schemes from raw data. In: *Joint Eurographics and IEEE VGTC Symposium on Visualization (EuroVis)*, *Computer Graphics Forum*, vol. 27, pp. 911–918 (2008)
13. Mayer, R., Roiger, A., Rauber, A.: Map-based interfaces for information management in large text collections. *Journal of Digital Information Management* 6(4), 294–302 (2008)
14. Plaisant, C., Rose, J., Yu, B., Auvil, L., Kirschenbaum, M.G., Smith, M.N., Clement, T., Lord, G.: Exploring erotics in emily dickinson’s correspondence with text mining and visual interfaces. In: *JCDL’06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pp. 141–150. ACM, New York (2006)
15. Poulet, F.: Towards Effective Visual Data Mining with Cooperative Approaches, pp. 389–406. Springer, Heidelberg (2008)
16. Rheingans, P., des Jardins, M.: Visualizing high-dimensional predictive model quality. In: *Proceedings of IEEE Visualization*, pp. 493–496 (2000)
17. Sabol, V., Kienreich, W., Muhr, M., Klieber, W., Granitzer, M.: Visual knowledge discovery in dynamic enterprise text repositories. In: *IV ’09: Proceedings of the 2009 13th International Conference Information Visualisation*, pp. 361–368. IEEE Computer Society, Washington (2009)
18. Seifert, C., Lex, E.: A novel visualization approach for data-mining-related classification. In: *Proceedings of the 13th International Conference on Information Visualisation (IV)*, July 2009, pp. 490–495. Wiley, Chichester (2009)
19. Seifert, C., Lex, E.: A visualization to investigate and give feedback to classifiers. Poster and Demo at Eurovis 2009 (June 2009) (unpublished)

20. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
21. Ware, M., Eibe, F., Holmes, G., Hall, M., Witten, I.H.: Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies* 55(3), 281–292 (2001)
22. Wong, P.C., Thomas, J.: Visual analytics. *IEEE Computer Graphics and Applications* 24, 20–21 (2004)
23. Zhao, Y., Karypis, G.: Evaluation of hierarchical clustering algorithms for document datasets. In: *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pp. 515–524. ACM Press, New York (2002)
24. Zhu, X.: Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin (2008)