Tjoa, Wagner (Eds.): Proceedings of the Dexa '08, Turin
19th International Conference on Database and Expert Systems Applications
ISBN 978-0-7695-3299-8, pp. 34-39, ©IEEE 2008.

Meta Analysis within Authorship Verification

Benno Stein Nedim Lipka Sven Meyer zu Eissen

Faculty of Media, Media Systems
Bauhaus University Weimar, Germany,
<first>.<last>@medien.uni-weimar.de

Abstract

In an authorship verification problem one is given writing examples from an author A, and one is asked to determine whether or not each text in fact was written by A. In a more general form of the authorship verification problem one is given a single document d only, and the question is whether or not d contains sections from other authors. The heart of authorship verification is the quantization of an author's writing style along with an outlier analysis to identify anomalies. Human readers are well-versed in detecting such spurious sections since they combine a highly-developed sense for wording with context-dependent meta knowledge in their analysis.

The intention of this paper is to compile an overview of the algorithmic building blocks for authorship verification. In particular, we introduce authorship verification problems as decision problems, discuss possibilities for the use of meta knowledge, and apply meta analysis to post-process unreliable style analysis results. Our meta analysis combines a confidence-based majority decision with the unmasking approach of Koppel and Schler. With this strategy we can improve the analysis quality in our experiments by 33% in terms of the F-measure.

Keywords *Authorship Verification, Plagiarism Analysis, Meta Learning*

1. Introduction

Authorship verification is a one-class classification problem. In a one-class classification problem one is given a target class for which a certain number of examples exist. Objects outside the target class are called outliers, and the one-class classification task is to tell apart outliers from target class members. Actually, the set of "outliers" can be much bigger than the target class, and an arbitrary number of outlier examples could be collected. Hence a one-class classification problem may look like a two-class discrimination problem; however, there is an important difference: members of the target class can be considered as representatives for their class, whereas one will not be able to compile a set of outliers that is representative for some kind of "non-target class". This fact is rooted in the enormous number and diversity of the non-target objects. Put another way, solving a one-class classification problem means to learn a concept (the concept of the target class) in the absence of discriminating features.[1]

Within authorship verification the target class is comprised of writing examples of a certain author A, whereas each piece of text written by another author B, $B \neq A$, is an outlier. In their excellent paper [10] Koppel and Schler give an illustrative discussion of authorship verification as a one-class classification problem. At the same place they introduce a new approach, called unmasking, to determine with a high probability whether a set of writing examples is a subset of the target class. Observe the term "set" in this connection: the unmasking approach does not solve the one-class classification problem for a single object but requires a batch of objects all of which must stem either from the target class or not (see details in Section 3).

1.1 Authorship Verification Problems

The complexity of an authorship verification problem can vary significantly, depending on the given constraints and assumptions. To organize existing research and the developed approaches we introduce, for the first time, three authorship verification problems—formulated as decision problems.

Problem. AVFIND
Given. A text d, allegedly written by author A.
Q. Does d contain sections written by an author B, $B \neq A$?

Problem. AVOUTLIER
Given. A set of texts $D = \{d_1, \ldots, d_n\}$, allegedly written by author A.
Q. Does D contain texts written by an author B, $B \neq A$?

[1] In rare cases, knowledge about outliers can be used to construct "representative counter examples" with respect to the target class. Then a standard discrimination approach can be applied.

Problem. AVBATCH
Given. Two sets of texts, $D_1 = \{d_{1_1}, \ldots, d_{1_k}\}$ and $D_2 = \{d_{2_1}, \ldots, d_{2_l}\}$, each of which written by a single author.
Q. Are the texts in D_1 and D_2 written by the same author?

Note that the problems can be transformed into each other, for example:

$$\text{AVFIND} \longrightarrow \text{AVOUTLIER} \longrightarrow \text{AVBATCH} \qquad (1)$$

Given a document d an AVFIND problem can be transformed into an AVOUTLIER problem by extracting suspicious sections from d. The AVOUTLIER problem in turn can be transformed into an AVBATCH problem by forming two sets D_1 and D_2 containing the suspicious and the non-suspicious sections respectively.

Note that the authorship verification problem AVFIND and intrinsic plagiarism analysis represent two sides of the same coin: the goal of intrinsic plagiarism analysis is to identify potentially plagiarized sections by analyzing a document with respect to changes in writing style [14].

1.2 Existing Research

Research related to authorship verification divides into the following areas: (*i*) models for the quantification of writing style—using classical measures for text complexity and grading level assessment [1, 7, 8, 6, 3, 4, 24] as well as author-specific stylometric analyses [18, 19, 11, 10, 9], (*ii*) technology for outlier analysis and machine learning [22, 23, 15, 12], and (*iii*) meta knowledge processing. Regarding the last area we refer to techniques for knowledge representation, deduction, and symbolic knowledge processing [17, 20].

1.3 Contributions

This paper deals with the solution of AVFIND. It overviews the involved algorithmic building blocks and discusses the possibilities for the use of meta analyses (Section 2). In particular we propose to solve AVFIND using transformation (1) : a document d is decomposed into sections $s_1, \ldots, s_n$, the sections are compared to the average writing style in d and labeled as outlier or not. If at least one s_i is labeled as an outlier, the answer to the AVOUTLIER problem could be "Yes" and consequently the answer to the AVFIND problem would be "Yes" as well. This corresponds to a minimum risk strategy. However, to gain further evidence for this hypothesis, a meta analysis is applied (Section 3): based on the solution of the AVOUTLIER problem an instance of AVBATCH is stated. Depending on the significance of its solution the hypothesis is accepted or rejected. In this connection the paper combines a confidence-based majority decision with an adapted version of the unmasking technology of Koppel and Schler [10].

2. Building Blocks for Authorship Verification

Transformation (1) shows connections of decreasing complexity: AVFIND is comprised of both a selection problem (finding suspicious sections) and an AVOUTLIER problem; likewise, the AVBATCH problem is a restricted variant of the AVOUTLIER problem since one has the additional knowledge that all elements of a batch are (or are not) outliers at the same time. I.e., AVFIND defines an all-encompassing authorship verification process. We organize this process into three stages:

1. A pre-analysis stage, where a knowledge-based "impurity" assessment may give us hints to find suspicious sections in a document d, and where a tailored decomposition strategy is chosen. These decisions in turn influence the construction of a model for style quantification.
2. A classification stage, where style outliers are identified with respect to the average writing style in d.
3. A post-processing stage, where the result of the classification stage is further analyzed with additional knowledge or technology. Main objective of this stage is the improvement of the analysis' overall precision and recall.

Table 1 organizes building blocks to operationalize these stages. Each column lists methods that can be applied, combined, or adapted in order to address a certain subtask in the entire authorship verification process. If these meta analyses happen in a skillful manner we may end up with an analysis process comparable to the power of a human reader; her/his salient strength is the integration of context-dependent meta knowledge in the analysis. The following subsections provide a comprehensive overview of places where meta knowledge can be operationalized.

2.1. Pre-Analysis Stage

Impurity Assessment. How likely is the fact that a document d contains a section from another author? We expect that the lengths, the places, and the entire fraction θ of such sections depend on particular document characteristics. Hence it makes sense to analyze the document type (paper, dissertation), its genre (novel, factual report, research, dictionary entry), but also the issuing institution (university, company, public service). Algorithmic means to reveal such information interpret document lengths, genres, and occurring named entities.

Decomposition Strategy. The simplest strategy is the decomposition of d into sections of equal length; in [14] the authors integrate an additional sentence detection. However, a more sensible interpretation of structural boundaries (chapters, paragraphs) is possible, which should consider

Pre-analysis			Classification	Post-processing	
Impurity assessment	**Decomposition strategy**	**Style model construction**	**Style outlier identification**	**Improvement at section level**	**Improvement at document level**
Document length analysis	Uniform length	Writer-specific: vocabulary richness	Two-class discriminant analysis	Citation analysis	Confidence-based majority decision
Genre Analysis	Structural boundaries	Writer-specific: complexity measures	One-class classifier: density estimation		Unmasking
Analysis of issuing institution	Text element boundaries	Reader-specific: grading level assessment	One-class classifier: boundary estimation		Batch means
	Topical boundaries	n-gram features	One-class classifier: reconstruction		Human inspection
		Language modeling			

Table 1. Building blocks of an authorship verification process. The first three columns contain pre-analysis methods, the fourth column comprises the classifier methods, which form the heart of each verification process, and the last two columns contain post-processing methods to improve the analysis quality.

special text elements like tables, formulas, footnotes, or quotations as well [16]. Though quite difficult, the detection of topical boundaries has a significant impact on the usefulness of a decomposition [2]. In [5] the authors even try to identify stylistic boundaries.

Style Model Construction. The decisions within the preceding steps must be considered within the style model construction: different stylometric features have different strengths but also pose different constraints on text length, text genre, or topic variation. In [14] connections of this type have been analyzed for the Flesch Kincaid Grade Level [4, 8], the Dale-Chall formula [3, 1], Yule's K [24], Honore's R [7], the Gunning Fog index [6], and the averaged word frequency class [13].

Applying a decomposition strategy to d yields a sequence of sections, $s_1, \ldots, s_n$, and the application of a style model yields a sequence of feature vectors $\mathbf{s}_1, \ldots, \mathbf{s}_n$. In the following classification stage the feature vectors are analyzed with respect to outliers.

2.2. Classification Stage

Recall that the identification of outliers among the $\mathbf{s}_i$ has to be solved solely on the basis of positive examples and therefore poses a one-class classification problem. Though one can reformulate the problem as a two-class discrimination problem by compiling a second class with some outliers, this is a bad advice. Usually, a tailored one-class classification approach should be applied; according to [22] such approaches fall into one of the following three classes:

(a) Density methods, which directly estimate the probability distributions of features for the target class. Outliers are assumed to be uniformly distributed, and Bayes' rule can be applied to separate outliers from the target class.

(b) Boundary methods, which avoid the estimation of the multi-dimensional density function but focus on the definition of a boundary around the set of target objects. The computation of the boundary is based on the distances between the objects in the target set.

(c) Reconstruction Methods. If we are given both an object's feature vector (the style model representation $\mathbf{s}$) as well as the original object (the section s), we may be able to reconstruct $\mathbf{s}$ from s as $\alpha(s)$, as well as to measure the reconstruction error $\alpha(s) \ominus \mathbf{s}$. It is assumed that α captures the domain theory underlying the target class, and the smaller the reconstruction error is the more likely s belongs to the target class.

Tax [22] investigates different representatives for these approaches: mixture of Gaussians, Parzen density, k-center, nearest neighbor, support vector data description, k-means, and self organizing maps. In particular, Tax provides meta knowledge to select among these classifiers by interpreting the presence of outliers, the scaling sensitivity, the number of free parameters, or the sample size.

2.3. Post-Processing Stage

We distinguish between methods that can be applied to a single section and methods that need the information of the entire document. Our overall objective is a confidence improvement of the analysis results obtained in the preceding stages.

Improvement at Section Level. A section may be correctly classified as a style outlier because it represents a citation. Such outliers must be excluded from an authorship verification. Due to their unique form of appearance, citations can be identified with a small number of heuristic rules.

Improvement at Document Level. Eventually, the result of the classification stage can only be improved by a *meta* learning approach. The idea is to use the solution of the

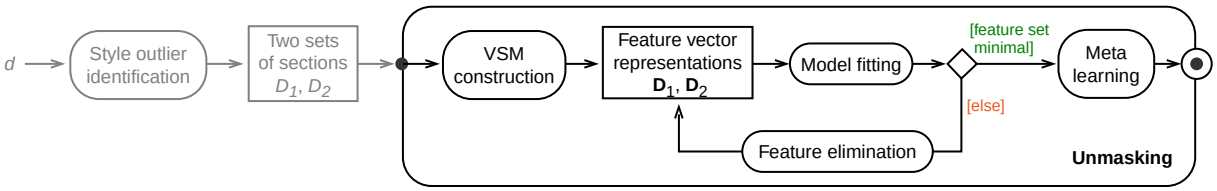

Figure 1. Unmasking: given are two sets D_1 and D_2 of outlier sections and target sections. Basic idea is to measure the separability of $\mathbf{D}_1$ versus $\mathbf{D}_2$ when the style model is successively impaired.

AVOUTLIER problem to form two sets D_1 (all sections labeled as outliers) and D_2 (all other sections), obtaining this way an instance of the AVBATCH problem; this idea was proposed for the first time by Stein and Meyer zu Eissen [21]. Possible meta learning approaches:

(a) Confidence-based majority decision. Based on a hypothesis for the impurity fraction θ of foreign text in a document d, one can assess an acceptance threshold τ for the number of outlier sections. The answer to the AVOUTLIER problem is "Yes" iff $|D_1| \geq \tau$.

(b) Koppel and Schler [10] developed the unmasking approach, which can be applied to solve AVBATCH. At heart, unmasking is a representative of what Tax terms "reconstruction method"; it measures the increase of a sequence of reconstruction errors, starting with a good reconstruction which then is more and more impaired. For two sets of texts from the same author the reconstruction error develops differently compared to the case where the two sets of texts are written by different authors.

(c) Batch means is a method that is applied within the analysis of simulation data in order to detect the end of a transient phase. For a sequence of values the variance development of the sample mean is measured while the sample size is successively increased. By processing the elements in D_1 and D_2 in a sorted manner, this idea can be adapted to solve instances of an AVBATCH problem.

The following section introduces unmasking in greater detail and reports on an evaluation.

3. Confidence Improvement by Post-processing

Given is an instance of the AVBATCH problem that was created as follows: a sufficiently large document d was decomposed into a number of sections that in turn were classified into two sets, D_1, D_2, assuming that all sections in D_2 belong to author A while all sections in D_1 belong to author B. By applying meta learning we now seek further evidence whether to accept the hypothesis $B \neq A$.

3.1. Unmasking

While the principle of a confidence-based majority decision is obvious, the unmasking approach requires a closer look to understand its rationale. At first, the sets D_1 and D_2 are represented under a reduced vector space model, designated as $\mathbf{D}_1$ and $\mathbf{D}_2$. As an initial feature set the 250 words with the highest (relative) frequency in $D_1 \cup D_2$ are chosen. Unmasking happens in the following steps (see Figure 1) :

1. *Model Fitting.* Training of a classifier that is able to separate $\mathbf{D}_1$ from $\mathbf{D}_2$. The authors in [10] implement a ten-fold cross-validation experiment using a linear kernel SVM to determine the achievable accuracy.
2. *Impairing.* Elimination of the most discriminative features with regard to the model obtained in Step 1; construction of new collections $\mathbf{D}_1$, $\mathbf{D}_2$, which now contain the impaired representations of the sections. [10] reports on convincing results by eliminating the six most discriminating features. Note, however, this heuristic depends on the section length which in turn depends on the length of d.
3. Go to Step 1 until the feature set is sufficiently reduced. Typically about 5-10 iterations are necessary.
4. *Meta Learning.* Analyze the degradation in the quality of the model fitting process: if after the last impairing step the sets $\mathbf{D}_1$ and $\mathbf{D}_2$ can still be separated with a small error, assume that d_1 and d_2 stem from different authors. Figure 2 shows a characteristic plot where unmasking was applied to short papers of 4-8 pages.

Rationale of unmasking: two sets of sections, D_1, D_2, constructed from two different documents d_1 and d_2 of the same author can be told apart easily if a vector space model (VSM) retrieval model is chosen. The VSM considers all words in $d_1 \cup d_2$, and hence it includes all kinds of open class and closed class word sets. If only the 250 most-frequent words are selected, a large fraction of them will be function words and stop words.[2] Among these 250 most-frequent words a small number does the major part of the discrimination job. These words capture topical differences,

[2] Function words and stop words are not disjunct sets: most function words in fact are stop words; however, the converse does not hold.

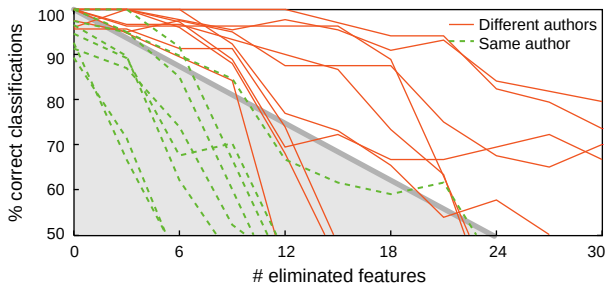

Figure 2. Unmasking at work: each line corresponds to a comparison of two papers, where a solid red (dashed green) line belongs to papers from two different authors (the same author).

differences that result from genre, purpose, or the like. By eliminating them, one approaches step by step the distinctive and subconscious manifestation of an author's writing style. After several iterations the remaining features are not powerful enough to discriminate two documents of the same author. By contrast, if d_1 and d_2 stem from two different authors, the remaining features will still quantify significant differences between the impaired representations $\mathbf{D}_1$ and $\mathbf{D}_2$ of the two sets of sections D_1 and D_2.

3.2. Experiments

The experiment design is oriented at Transformation (1): given a document d we solve AVFIND ("Does d contain text from a foreign author?") by formulating an instance of AVOUTLIER. For this purpose d is decomposed into sections, which, in a first step, are classified as being either style-conform or not. If at least one section is labeled as a style outlier, the answer to AVFIND *under a minimum risk strategy* is "Yes". In our experiments we apply the more sensible strategy introduced before and post-process the AVOUTLIER results by constructing and solving an AVBATCH problem.

Our test corpus contains scientific documents written in German. Basis of the corpus are dissertation theses and habilitation theses from the following fields: philosophy, psychology, sociology, medical science, historical science, and law. From the original theses all explicitly declared citations were removed, and clippings of about 10 000 words were extracted. These clippings represent the "clean" documents written by a single author; 10% of the clean documents were used to construct impure documents by inserting between 4 and 8 sections of 500 words from foreign authors.

The experiment setup as well as the problem-specific parameters for AVFIND, AVOUTLIER, and AVBATCH are as follows:

AVFIND

portion of impure documents in the corpus:	0.10
average document lengths (words):	10 000
impurity fraction θ per impure document:	0.20…0.40

AVOUTLIER

impure sections per impure document:	4…8
size of section (words):	500
training set size / test set size:	10-fold cross validation
classifier type:	logistic regression
style model features:	word class distribution, text complexity measures, vowel-consonant trigrams

AVBATCH (majority)

upper bound $\tau_=$ for $B = A$:	$\lvert D_1\rvert \leq 4$
lower bound $\tau_\neq$ for $B \neq A$:	$\lvert D_1\rvert \geq 6$
uncertainty domain:	otherwise

AVBATCH (unmasking)

model fitting approach:	support vector machine
number of iterations:	10

To solve instances of AVOUTLIER a one-class classifier is required. For this purpose we employed the most powerful style features that we found in previous work [14, 21]. Our trained classifier achieves a recall of 0.80 for both the class of outlier sections and the class of target sections. [3]

The recall of outlier sections (sections presumably from a foreign author B) tells us something about the maximum possible identification rate of a foreign author's text; similarly, *1 minus the recall* of target sections (sections presumably from author A) defines the probability of asserting a wrong claim with respect to a style change. Consider in this connection the class imbalance between outliers and targets: since only each tenth document is impure, for instance with a fraction of $\theta = 0.2$, the ratio between outliers and targets is 1:49, assuming 20 sections per document. I.e., under a minimum risk strategy where one answers an AVOUTLIER problem with "Yes" if at least one section is labeled as outlier ($B \neq A$), nearly all documents are misleadingly claimed as impure.

By post-processing the AVOUTLIER results we gain more confidence with respect to the AVFIND problem and accept or reject the result obtained under the minimum risk strategy. Post-processing happens in a combined fashion: using the majority decision approach we answer AVFIND with "No" (d does not contain sections from a foreign author) if the number of classified outliers, $|D_1|$, is below the bound $\tau_=$. Likewise, we answer AVFIND with "Yes" (d contains section from a foreign author) if the number of classified outliers, $|D_1|$, is above the bound $\tau_\neq$. If $|D_1|$ is in the uncertainty domain, $\tau_= < |D_1| < \tau_\neq$, we solve AVBATCH with the unmasking approach. Table 2 comprises the achieved classification results.

[3] The recall of class x is defined as $P(h(\mathbf{s})=x|x)$, the probability that a classifier h labels a section s as x, given the fact that s belongs to class x.

	Classification			Post-processing					
	AVOUTLIER			AVBATCH			AVBATCH		
Impurity	**(minimum risk)**			**(majority)**			**(unmasking)**		
θ	*prec*	*rec*	F	*prec*	*rec*	F	*prec*	*rec*	F
0.20	0.12	1.00	0.56	0.71	0.83	0.77	0.73	0.90	0.82
0.30	0.20	1.00	0.60	1.00	0.56	0.78	1.00	0.93	0.97
0.40	0.18	1.00	0.59	1.00	0.83	0.92	1.00	0.87	0.94

Table 2. Classification results: the AVOUTLIER problem under the minimum risk strategy (column 2-4), the related AVBATCH problem under a majority decision approach (column 5-7), and under unmasking (column 8-10).

The optimum bounds for the majority decision approach, $\tau_=$ and $\tau_\neq$, are computed within a supervised learning stage, using the same training set that had been used to learn the one-class classifier for the AVOUTLIER problem.

4. Discussion

The improvements achieved by the meta analysis within the post-processing stage are substantial (see Table 2). We would like to point out that the approach of Koppel and Schler unfolds its power especially if an impure document is mistakenly classified as a document from a single author. The case that $|D_1|$ is in the uncertainty domain happens in 3% of all AVBATCH instances, and in 30% of all impure AVBATCH instances.

Finally, observe the following tradeoff: with increasing θ the solution of AVOUTLIER becomes more difficult, but the solution of AVBATCH becomes simpler. Rationale for the former is that an increasing θ masks possible style deviations from a document's averaged writing style model. Rationale for the latter is the availability of more sample texts to apply unmasking.

References

[1] J. Chall and E. Dale. *Readability Revisited: The new Dale-Chall Readability Formula*. Brookline Books, 1995.

[2] F. Choi. Advances in domain independent linear text segmentation. *Proc. of the first conf. on North American chapter of the Association for Computational Linguistics*. Morgan Kaufmann, 2000

[3] E. Dale and J. Chall. A formula for predicting readability. *Educ. Res. Bull.*. 1948.

[4] R. Flesch. A new readability yardstick. *J. of Applied Psychology*. 1948.

[5] N. Graham, G. Hirst, and B. Marthi. Segmenting a document by stylistic character. *Natural Language Engineering*. 2005.

[6] R. Gunning. *The Technique of Clear Writing*. McGraw-Hill. 1952.

[7] A. Honore. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*. 1979.

[8] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom. Derivation of new readability formulas for navy enlisted personnel. *Research Branch Report 8 75 Millington TN: Naval Technical Training US Naval Air Station*. 1975.

[9] M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. *Proc. of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*. 2003.

[10] M. Koppel and J. Schler. Authorship Verification as a One-Class Classification Problem. *Proc. of the 21st Int. Conf. on Machine Learning*. ACM, 2004.

[11] M. Koppel, J. Schler, S. Argamon, and E. Messeri. Authorship attribution with thousands of candidate authors. *Proc. of the 29th annual int. ACM SIGIR conf. on Research and development in information retrieval*. ACM, 2006.

[12] L. Manevitz and M. Yousef. One-Class SVMs for Document Classification. *J. of Machine Learning Research*. 2001.

[13] S. Meyer zu Eißen and B. Stein. Genre Classification of Web Pages: User Study and Feasibility Analysis. *KI 2004: Advances in Artificial Intelligence*. Springer, 2004.

[14] S. Meyer zu Eissen and B. Stein. Intrinsic plagiarism detection. *Proc. of the European Conf. on Information Retrieval (ECIR 2006)*. Springer, 2006.

[15] G. Ratsch, S. Mika, B. Scholkopf, and K.-R. Muller. Constructing Boosting Algorithms from SVMs: An Application to One-Class Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2002.

[16] J. Reynar. *Topic Segmentation: Algorithms and Applications*. PhD thesis, University of Pennsylvania, 1998.

[17] S. Russel and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, 1995.

[18] E. Stamatatos. Author Identification Using Imbalanced and Limited Training Texts. *18th Int. Conf. on Database and Expert Systems Applications (DEXA 07)*. 2007.

[19] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 2001.

[20] M. Stefik. *Introduction to Knowledge Systems*. Morgan Kaufmann, 1995.

[21] B. Stein and S. Meyer zu Eissen. Intrinsic Plagiarism Analysis with Meta Learning. *SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07)*. 2007.

[22] D. Tax. *One-Class Classification*. PhD thesis, Technische Universiteit Delft, 2001.

[23] D. Tax and R. Duin. Combining One-Class Classifiers. *Proc. of the Second Int. Workshop on Multiple Classifier Systems*. Springer, 2001.

[24] G. Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1944.