Using Syntactic Information to Identify Plagiarism

Özlem Uzuner, Boris Katz, and Thade Nahnsen
Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139
ozlem,boris,tnahnsen@csail.mit.edu

Abstract

Using keyword overlaps to identify plagiarism can result in many false negatives and positives: substitution of synonyms for each other reduces the similarity between works, making it difficult to recognize plagiarism; overlap in ambiguous keywords can falsely inflate the similarity of works that are in fact different in content. Plagiarism detection based on verbatim similarity of works can be rendered ineffective when works are paraphrased even in superficial and immaterial ways. Considering linguistic information related to creative aspects of writing can improve identification of plagiarism by adding a crucial dimension to evaluation of similarity: documents that share linguistic elements in addition to content are more likely to be copied from each other. In this paper, we present a set of low-level syntactic structures that capture creative aspects of writing and show that information about linguistic similarities of works improves recognition of plagiarism (over tfidf-weighted keywords alone) when combined with similarity measurements based on tfidf-weighted keywords.

1 Introduction

To plagiarize is "to steal and pass off (the ideas or words of another) as one's own; [to] use (another's production) without crediting the source; [or] to commit literary theft [by] presenting as new and original an idea or product derived from an existing source".[1] Plagiarism is frequently encountered in academic settings. According to turnitin.com, a 2001 survey of 4500 high school students revealed that "15% [of students] had submitted a paper obtained in large part from a term paper mill or website". Increased rate of plagiarism hurts quality of education received by students; facilitating recognition of plagiarism can help teachers control this damage.

To facilitate recognition of plagiarism, in the recent years many commercial and academic products have been developed. Most of these approaches identify verbatim plagiarism[2] and can fail when works are paraphrased. To recognize plagiarism in paraphrased works, we need to capture similarities that go beyond keywords and verbatim overlaps. Two works that exhibit similarity both in their conceptual content (as indicated by keywords) and in their expression of this content should be considered more similar than two works that are similar only in content. In this context, *content* refers to the story or the information; *expression* refers to the linguistic choices of authors used in presenting the content, i.e., creative elements of writing, such as whether authors tend toward passive or active voice, whether they prefer complex sentences with embedded clauses or simple sentences with independent clauses, as well as combinations of such choices.

Linguistic information can be a source of power for measuring similarity between works based on

[1] www.webster.com
[2] www.turnitin.com

Proceedings of the 2nd Workshop on Building Educational Applications Using NLP,
pages 37–44, Ann Arbor, June 2005.

their expression of content. In this paper, we use linguistic information related to the creative aspects of writing to improve recognition of paraphrased documents as a first step towards plagiarism detection. To identify a set of features that relate to the linguistic choices of authors, we rely on different syntactic expressions of the same content. After identifying the relevant features (which we call *syntactic elements of expression*), we rely on patterns in the use of these features to recognize paraphrases of works.

In the absence of real-life plagiarism data, in this paper, we use a corpus of parallel translations of novels as surrogate for plagiarism data. Translations of *titles*, i.e., original works, into English by different people provide us with *books* that are paraphrases of the same content. We use these paraphrases to automatically identify:

1. Titles even when they are paraphrased, and

2. Pairs of book chapters that are paraphrases of each other.

Our first experiment shows that syntactic elements of expression outperform all baselines in recognizing titles even when they are paraphrased, providing a way of recognizing copies of works based on the similarities in their expression of content. Our second experiment shows that similarity measurements based on the combination of tfidf-weighted keywords and syntactic elements of expression outperform the weighted keywords in recognizing pairs of book chapters that are paraphrases of each other.

2 Related Work

We define expression as "the linguistic choices of authors in presenting a particular content" (Uzuner, 2005; Uzuner and Katz, 2005). Linguistic similarity between works has been studied in the text classification literature for identifying the style of an author. However, it is important to differentiate expression from style. Style refers to the *linguistic elements that, independently of content, persist over the works* of an author and has been widely studied in authorship attribution. Expression involves the *linguistic elements that relate to how an author phrases particular content* and can be used to identify potential copyright infringement or plagiarism. Similarities in the expression of similar content in two different works signal potential copying. We hypothesize that syntax plays a role in capturing expression of content. Our approach to recognizing paraphrased works is based on phrase structure of sentences in general, and structure of verb phrases in particular.

Most approaches to similarity detection use computationally cheap but linguistically less informed features (Peng and Hengartner, 2002; Sichel, 1974; Williams, 1975) such as keywords, function words, word lengths, and sentence lengths; approaches that include deeper linguistic information, such as syntactic information, usually incur significant computational costs (Uzuner et al., 2004). Our approach identifies useful linguistic information without incurring the computational cost of full text parsing; it uses context-free grammars to perform high-level syntactic analysis of part-of-speech tagged text (Brill, 1992). It turns out that such a level of analysis is sufficient to capture syntactic information related to creative aspects of writing; this in turn helps improve recognition of paraphrased documents. The results presented here show that extraction of useful linguistic information for text classification purposes does not have to be computationally prohibitively expensive, and that despite the tradeoff between the accuracy of features and computational efficiency, we can extract linguistically-informed features without full parsing.

3 Identifying Creative Aspects of Writing

In this paper, we first identify linguistic elements of expression and then study patterns in the use of these elements to recognize a work even when it is paraphrased. Translated literary works provide examples of linguistic elements that differ in expression but convey similar content. These works provide insight into the linguistic elements that capture expression. For example, consider the following semantically equivalent excerpts from three different translations of *Madame Bovary* by Gustave Flaubert.

> Excerpt 1: "Now Emma would often take it into her head to write him during the day. Through her window she would signal to Justin, and he would whip off his apron and fly to la huchette. And when Rodolphe arrived in response to her summons, it was to hear that she was miserable, that her husband was odious, that her life was a torment." (Translated by Unknown1.)

> Excerpt 2: "Often, even in the middle of the day, Emma suddenly wrote to him, then from the window made a sign to Justin, who, taking his apron off, quickly ran to la huchette. Rodolphe would come; she had sent for him to tell him that she was bored, that her husband was odious, her life frightful." (Translated by Aveling.)
>
> Excerpt 3: "Often, in the middle of the day, Emma would take up a pen and write to him. Then she would beckon across to Justin, who would off with his apron in an instant and fly away with the letter to la huchette. And Rodolphe would come. She wanted to tell him that life was a burden to her, that she could not endure her husband and that things were unbearable." (Translated by Unknown2.)

Inspired by syntactic differences displayed in such parallel translations, we identified a novel set of syntactic features that relate to how people convey content.

3.1 Syntactic Elements of Expression

We hypothesize that given particular content, authors choose from a set of semantically equivalent syntactic constructs to express this content. To paraphrase a work without changing content, people try to interchange semantically equivalent syntactic constructs; patterns in the use of various syntactic constructs can be sufficient to indicate copying.

Our observations of the particular expressive choices of authors in a corpus of parallel translations led us to define syntactic elements of expression in terms of sentence-initial and -final phrase structures, semantic classes and argument structures of verb phrases, and syntactic classes of verb phrases.

3.1.1 Sentence-initial and -final phrase structures

The order of phrases in a sentence can shift the emphasis of a sentence, can attract attention to particular pieces of information and can be used as an expressive tool.

1 (a) Martha can finally put some money in the bank.
(b) Martha can put some money in the bank, finally.
(c) Finally, Martha can put some money in the bank.

2 (a) Martha put some money in the bank on Friday.
(b) On Friday, Martha put some money in the bank.
(c) Some money *is what* Martha put in the bank on Friday.
(d) In the bank *is where* Martha put some money on Friday.

The result of such expressive changes affect the distributions of various phrase types in sentence-initial and -final positions; studying these distributions can help us capture some elements of expression. Despite its inability to detect the structural changes that do not affect the sentence-initial and -final phrase types, this approach captures some of the phrase-level expressive differences between semantically equivalent content; it also captures different sentential structures, including question constructs, imperatives, and coordinating and subordinating conjuncts.

3.1.2 Semantic Classes of Verbs

Levin (1993) observed that verbs that exhibit similar syntactic behavior are also related semantically. Based on this observation, she sorted 3024 verbs into 49 high-level semantic classes. Verbs of "sending and carrying", such as `convey`, `deliver`, `move`, `roll`, `bring`, `carry`, `shuttle`, and `wire`, for example, are collected under this semantic class and can be further broken down into five semantically coherent lower-level classes which include "drive verbs", "carry verbs", "bring and take verbs", "slide verbs", and "send verbs". Each of these lower-level classes represents a group of verbs that have similarities both in semantics and in syntactic behavior, i.e., they can grammatically undergo similar syntactic alternations. For example, "send verbs" can be seen in the following alternations (Levin, 1993):

1. **Base Form**
   - Nora sent the book to Peter.
   - NP + V + NP + PP.
2. **Dative Alternation**
   - Nora sent Peter the book.
   - NP + V + NP + NP.

Semantics of verbs in general, and Levin's verb classes in particular, have previously been used for evaluating content and genre similarity (Hatzivassiloglou et al., 1999). In addition, similar semantic classes of verbs were used in natural language processing applications: START was the first natural language question answering system to use such verb classes (Katz and Levin, 1988). We use

Levin's semantic verb classes to describe the expression of an author in a particular work. We assume that semantically similar verbs are often used in semantically similar syntactic alternations; we describe part of an author's expression in a particular work in terms of the semantic classes of verbs she uses and the particular argument structures, e.g., NP + V + NP + PP, she prefers for them. As many verbs belong to multiple semantic classes, to capture the dominant semantic verb classes in each document we credit all semantic classes of all observed verbs. We extract the argument structures from part of speech tagged text, using context-free grammars (Uzuner, 2005).

3.1.3 Syntactic Classes of Verbs

Levin's verb classes include exclusively "non-embedding verbs", i.e., verbs that do not take clausal arguments, and need to be supplemented by classes of "embedding verbs" that do take such arguments. Alexander and Kunz (1964) identified syntactic classes of embedding verbs, collected a comprehensive set of verbs for each class, and described the identified verb classes with formulae written in terms of phrasal and clausal elements, such as verb phrase heads (Vh), participial phrases (Partcp.), infinitive phrases (Inf.), indicative clauses (IS), and subjunctives (Subjunct.). We used 29 of the more frequent embedding verb classes and identified their distributions in different works. Examples of these verb classes are shown in Table 1. Further examples can be found in (Uzuner, 2005; Uzuner and Katz, 2005).

Syntactic Formula	Example
NP + Vh + NP + from + Partcp.	The belt kept him from dying.
NP + Vh + that + IS	He admitted that he was guilty.
NP + Vh + that + Subjunct.	I request that she go alone.
NP + Vh + to + Inf.	My father wanted to travel.
NP + Vh + wh + IS	He asked if they were alone.
NP + pass. + Partcp.	He was seen stealing.

Table 1: Sample syntactic formulae and examples of embedding verb classes.

We study the syntax of embedding verbs by identifying their syntactic class and the structure of their observed embedded arguments. After identifying syntactic and semantic characteristics of verb phrases, we combine these features to create further elements of expression, e.g., syntactic classes of embedding verbs and the classes of semantic non-embedding verbs they co-occur with.

4 Evaluation

We tested sentence-initial and -final phrase structures, semantic and syntactic classes of verbs, and structure of verb arguments, i.e., syntactic elements of expression, in paraphrase recognition and in plagiarism detection in two ways:

- Recognizing titles even when they are paraphrased, and
- Recognizing pairs of book chapters that are paraphrases of each other.

For our experiments, we split books into chapters, extracted all relevant features from each chapter, and normalized them by the length of the chapter.

4.1 Recognizing Titles

Frequently, people paraphrase parts of rather than complete works. For example, they may paraphrase chapters or paragraphs from a work rather than the whole work. We tested the effectiveness of our features on recognizing paraphrased components of works by focusing on chapter-level excerpts (smaller components than chapters have very sparse vectors given our sentence-level features and will be the foci of future research) and using boosted decision trees (Witten and Frank, 2000).

Our goal was to recognize chapters from the *titles* in our corpus even when some *titles* were paraphrased into multiple *books*; in this context, titles are original works and paraphrased books are translations of these titles. For this, we assumed the existence of one legitimate book from each title. We used this book to train a model that captured the syntactic elements of expression used in this title. We used the remaining paraphrases of the title (i.e., the remaining books paraphrasing the title) as the test set—these paraphrases are considered to be plagiarized copies and should be identified as such given the model for the title.

4.1.1 Data

Real life plagiarism data is difficult to obtain. However, English translations of foreign titles exist and can be obtained relatively easily. Titles that have been translated on different occasions by different translators and that have multiple translations provide us with examples of books that paraphrase the same content and serve as our surrogate for plagiarism data.

To evaluate syntactic elements of expression on recognizing paraphrased chapters from titles, we compared the performance of these features with tfidf-weighted keywords on a 45-way classification task. The corpus used for this experiment included 49 books from 45 titles. Of the 45 titles, 3 were paraphrased into a total of 7 books (3 books paraphrased the title *Madame Bovary*, 2 books paraphrased *20000 Leagues*, and 2 books paraphrased *The Kreutzer Sonata*). The remaining titles included works from J. Austen (1775-1817), C. Dickens (1812-1870), F. Dostoyevski (1821-1881), A. Doyle (1859-1887), G. Eliot (1819-1880), G. Flaubert (1821-1880), T. Hardy (1840-1928), V. Hugo (1802-1885), W. Irving (1789-1859), J. London (1876-1916), W. M. Thackeray (1811-1863), L. Tolstoy (1828-1910), I. Turgenev (1818-1883), M. Twain (1835-1910), and J. Verne (1828-1905).

4.1.2 Baseline Features

The task described in this section focuses on recognizing paraphrases of works based on the way they are written. Given the focus of authorship attribution literature on "the way people write", to evaluate the syntactic elements of expression on recognizing paraphrased chapters of a work, we compared these features against features frequently used in authorship attribution as well as features used in content recognition.

Tfidf-weighted Keywords: Keywords, i.e., content words, are frequently used in content-based text classification and constitute one of our baselines.

Function Words: In studies of authorship attribution, many researchers have taken advantage of the differences in the way authors use function words (Mosteller and Wallace, 1963; Peng and Hengartner, 2002). In our studies, we used a set of 506 function words (Uzuner, 2005).

Distributions of Word Lengths and Sentence Lengths: Distributions of word lengths and sentence lengths have been used in the literature for authorship attribution (Mendenhall, 1887; Williams, 1975; Holmes, 1994). We include these features in our sets of baselines along with information about means and standard deviations of sentence lengths (Holmes, 1994).

Baseline Linguistic Features: Sets of surface, syntactic, and semantic features have been found to be useful for authorship attribution and have been adopted here as baseline features. These features included: the number of words and the number of sentences in the document; type–token ratio; average and standard deviation of the lengths of words (in characters) and of the lengths of sentences (in words) in the document; frequencies of declarative sentences, interrogatives, imperatives, and fragmental sentences; frequencies of active voice sentences, be-passives and get-passives; frequencies of 's-genitives, of-genitives and of phrases that lack genitives; frequency of overt negations, e.g., "not", "no", etc.; and frequency of uncertainty markers, e.g., "could", "possibly", etc.

4.1.3 Experiment

To recognize chapters from the *titles* in our corpus even when some *titles* were paraphrased into multiple *books*, we randomly selected 40–50 chapters from each title. We used 60% of the selected chapters from each title for training and the remaining 40% for testing. For paraphrased titles, we selected training chapters from one of the paraphrases and testing chapters from the remaining paraphrases. We repeated this experiment three times; at each round, a different paraphrase was chosen for training and the rest were used for testing.

Our results show that, on average, syntactic elements of expression accurately recognized components of titles 73% of the time and significantly outperformed all baselines[3] (see middle column in Table 2).[4]

[3]The tfidf-weighted keywords used in this experiment do not include proper nouns. These words are unique to each title and can be easily replaced without changing content or expression in order to trick a plagiarism detection system that would rely on proper nouns.

[4]For the corpora used in this paper, a difference of 4% or more is statistically significant with $\alpha = 0.05$.

Feature Set	Avg. accuracy (complete corpus)	Avg. accuracy (paraphrases) only
Syntactic elements of expression	73%	95%
Function words	53%	34%
Tfidf-weighted keywords	47%	38%
Baseline linguistic	40%	67%
Dist. of word length	18%	54%
Dist. of sentence length	12%	17%

Table 2: Classification results (on the test set) for recognizing titles in the corpus even when some titles are paraphrased (middle column) and classification results only on the paraphrased titles (right column). In either case, random chance would recognize a paraphrased title 2% of the time.

The right column in Table 2 shows that the syntactic elements of expression accurately recognized on average 95% of the chapters taken from paraphrased titles. This finding implies that some of our elements of expression are common to books that are derived from the same title. This commonality could be due to the similarity of their content or due to the underlying expression of the original author.

4.2 Recognizing Pairs of Paraphrased Chapters

Experiments in Section 4.1 show that we can use syntactic elements of expression to recognize titles and their components based on the way they are written even when some works are paraphrased. In this section, our goal is to identify pairs of chapters that paraphrase the same content, i.e., chapter 1 of translation 1 of *Madame Bovary* and chapter 1 of translation 2 of *Madame Bovary*. For this evaluation, we used a similar approach to that presented by Nahnsen et al. (2005).

4.2.1 Data

Our data for this experiment included 47 chapters from each of two translations of *20000 Leagues under the Sea* (Verne), 35 chapters from each of 3 translations of *Madame Bovary* (Flaubert), 28 chapters from each of two translations of *The Kreutzer Sonata* (Tolstoy), and 365 chapters from each of 2 translations of *War and Peace* (Tolstoy). Pairing up the chapters from these titles provided us with more than 1,000,000 chapter pairs, of which approximately 1080 were paraphrases of each other.[5]

4.2.2 Experiment

For experiments on finding pairwise matches, we used similarity of vectors of tfidf-weighted keywords;[6] and the multiplicative combination of the similarity of vectors of tfidf-weighed keywords of works with the similarity of vectors of syntactic elements of expression of these works. We used cosine to evaluate the similarity of the vectors of works. We omitted the remaining baseline features from this experiment—they are features that are common to majority of the chapters from each book, they do not relate to the task of finding pairs of chapters that could be paraphrases of each other.

We ranked all chapter pairs in the corpus based on their similarity. From this ranked list, we identified the top n most similar pairs and predicted that they are paraphrases of each other. We evaluated our methods with precision, recall, and f-measure.[7]

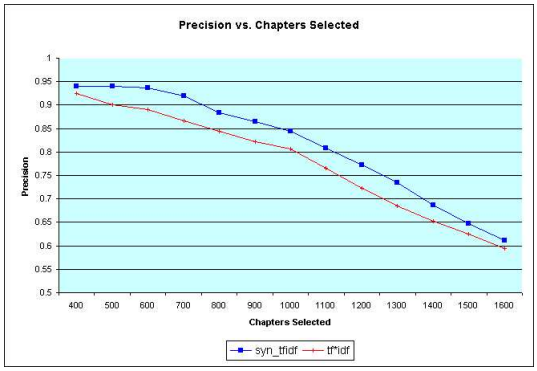

Figure 1: Precision.

Figures 1, 2, and 3 show that syntactic elements of expression improve the performance of tfidf-weighted keywords in recognizing pairs of paraphrased chapters significantly in terms of precision, recall, and f-measure for all n; in all of these figures, the blue line marked *syn_tfidf* represents the performance of tfidf-weighted keywords enhanced with

[5] Note that this number double-counts the paraphrased pairs; however, this fact is immaterial for our discussion.

[6] In this experiment, proper nouns are included in the weighted keywords.

[7] The ground truth marks only the same chapter from two different translations of the same title as similar, i.e., chapter x of translation 1 of *Madame Bovary* and chapter y of translation 2 of *Madame Bovary* are similar only when $x = y$.

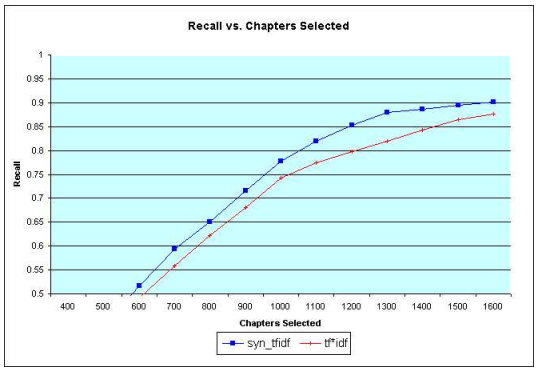

Figure 2: Recall.

syntactic elements of expression. More specifically, the peak f-measure for tfidf-weighted keywords is approximately 0.77 without contribution from syntactic elements of expression. Adding information about similarity of syntactic features to cosine similarity of tfidf-weighted keywords boosts peak f-measure value to approximately 0.82.[8] Although the f-measure of both representations degrade when $n > 1100$, this degradation is an artifact of the evaluation metric: the corpus includes only 1080 similar pairs, at $n > 1100$, recall is very close to 1, and therefore increasing *n* hurts overall performance.

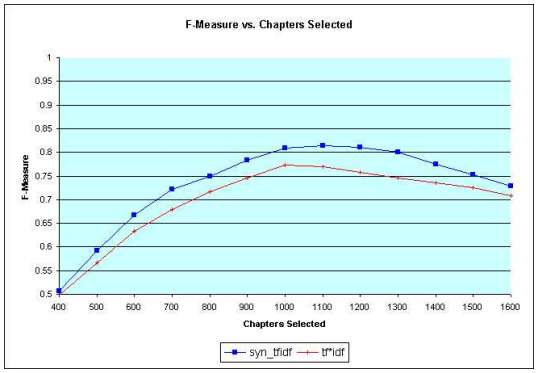

Figure 3: F-measure.

5 Conclusion

Plagiarism is a problem at all levels of education. Increased availability of digital versions of works makes it easier to plagiarize others' work and the large volumes of information available on the web makes it difficult to identify cases of plagiarism. To identify plagiarism even when works are paraphrased, we propose studying the use of particular syntactic constructs as well as keywords in documents.

This paper shows that syntactic information can help recognize works based on the way they are written. Syntactic elements of expression that focus on the changes in the phrase structure of works help identify paraphrased components of a title. The same features help improve identification of pairs of chapters that are paraphrases of each other, despite the content these chapters share with the rest of the chapters taken from the same title. The results presented in this paper are based on experiments that use translated novels as surrogate for plagiarism data. Our future work will extend our study to real life plagiarism data.

[8]The difference is statistically significant at $\alpha = 0.05$.

6 Acknowledgements

The authors would like to thank Sue Felshin for her insightful comments. This work is supported in part by the Advanced Research and Development Activity as part of the AQUAINT research program.

References

D. Alexander and W. J. Kunz. 1964. Some classes of verbs in English. In *Linguistics Research Project*. Indiana University, June.

E. Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*.

V. Hatzivassiloglou, J. Klavans, and E. Eskin. 1999. Detecting similarity by applying learning over indicators. In *Proceedings of the 37th Annual Meeting of the ACL*.

D. I. Holmes. 1994. Authorship attribution. *Computers and the Humanities*, 28.

B. Katz and B. Levin. 1988. Exploiting lexical regularities in designing natural language systems. In *Proceedings of the 12th Int'l Conference on Computational Linguistics (COLING '88)*.

B. Levin. 1993. *English Verb Classes and Alternations. A Preliminary Investigation*. University of Chicago Press.

T. C. Mendenhall. 1887. Characteristic curves of composition. *Science*, 11.

F. Mosteller and D. L. Wallace. 1963. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302).

T. Nahnsen, Ö. Uzuner, and B. Katz. 2005. Lexical chains and sliding locality windows in content-based text similarity detection. *CSAIL Memo*, AIM-2005-017.

R. D. Peng and H. Hengartner. 2002. Quantitative analysis of literary styles. *The American Statistician*, 56(3).

H. S. Sichel. 1974. On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society (A)*, 137.

Ö. Uzuner and B. Katz. 2005. Capturing expression using linguistic information. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*.

Ö. Uzuner, R. Davis, and B. Katz. 2004. Using empirical methods for evaluating expression and content similarity. In *Proceedings of the 37th Hawaiian International Conference on System Sciences (HICSS-37)*. IEEE Computer Society.

Ö. Uzuner. 2005. *Identifying Expression Fingerprints Using Linguistic Information*. Ph.D. thesis, Massachusetts Institute of Technology.

C. B. Williams. 1975. Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. *Biometrika*, 62(1).

I. H. Witten and E. Frank. 2000. *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann, San Francisco.