A Framework for Authorship Identification in the Internet Environment

Jan Rygl, Aleš Horák

NLP Centre
Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xrygl,hales}@fi.muni.cz

Abstract. Misuse of anonymous online communication for illegal purposes has become a major concern [2,12]. In this paper, we present a framework named *ART* (*Authorship Recognition Tool*), that is designed to minimize manual procedures and maximize the efficiency of authorship identification based on the content of Internet electronic documents. The framework covers the phases of document retrieval and database document management. *ART* provides implementations of efficient authorship identification algorithm and authorship similarity algorithm including the possibility to obtain extra data for learning and tests. The framework also determines whether or not different author's identities are interlinked.

The authorship is analysed by machine learning and natural language processing methods. Technical information such as IP address is considered only as an optional attribute for the machine learning because it can be easily forged or devalued if the author communicates from public places or through proxy servers. The decision which algorithm to use for determining the authorship of an anonymous document depends on the documents' language.

Key words: authorship identification, authorship similarity

1 Introduction

Recent extremist actions in the "civilized" parts of the world call for more efficient techniques for crime prevention. Relatively new medium for the communication of extremist groups and even individuals (when expressing their thoughts to the public) is the public Internet. This medium offers to the authors the possibility to publish very fast to a large audience with techniques available to remain anonymous. One way to defend the public against this anonymity is based on computational linguistic methods of forensic authorship analysis [1,9].

In the following text, we describe the design and implementation of a new framework, named *ART* (*Authorship Recognition Tool*), for effective solution of

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2011, pp. 117–124, 2011. © Tribun EU 2011

authorship identification in the online environment. The main task of the *ART* framework can be defined as:

Problem 1. Let us have an anonymous document *D*. The task is to identify the author of the document and find all other documents of the same author in the public Internet.

What most of currently prevailing techniques have in common is that they work with electronic communication from the crime, terrorism and extremism environment [5,10,11,12]. These studies are aimed on the identification methods, hence online messages are collected manually [10,11].

In this work a new approach is presented. The proposed framework is unique because it does not focus on various techniques of determining the authorship of documents, but it analyses the problem and offers solutions to many technical difficulties related to the authorship attribution in the Internet environment.

2 Framework for the Online Authorship Identification

The given problem, as we have specified in the introduction, is very complex, therefore it is necessary to decompose it into smaller, clearly defined tasks. Firstly, the scenario of a fully functional system is described. It works in 4 phases (as shown in Table 1):

1. In the first phase, the set of possible authors is restricted according to manually annotated domains' themes – if an unknown document from a Czech extremist website is analysed, only authors from this and other Czech extremist websites are taken into account. This step is important because the accuracy of authorship identification decreases with growing number of potential authors. Filtering out improbable authors increases the success rate of authorship detection algorithms at the cost of manual category tagging of selected web domains.
2. After narrowing the set of possible authors, each author's profile is compared to the unknown document by machine learning (comparison of several algorithms is available in [7]) and natural language processing methods (e.g. delta score [3], punctuation statistics [4,6]). The unknown document is associated with the author whose profile is the most similar to the document.
3. In the next phase, all web domains related to the theme selected in the first step are analysed. Their content is downloaded to the database that contains pairs of preprocessed documents and an author identity (a pseudonym or technical parameters can serve as the identity). Acquired documents are divided into groups according to their authorship. After the document clustering, groups are unified if their similarity exceeds a pre-set limit (their author published under more identities).
4. In the last phase, all documents of the guessed author are returned and new information are stored into the database.

Table 1. Process of the Authorship Identification

action	description

`Extract document's theme`
`Extract domains' themes`
`Select domains with related themes`
	(manually)
	(manually)
	section 3

`Compare known authors to the document`
`Select the most similar author`
	section 4

`Analyze documents from selected domains`
`Cluster documents according to their authorship`
	section 5
	section 6

`Return the cluster of the selected author`	

Each phase is described in more detail in the following sections, including the semi-automatic document downloading from Internet.

3 Restricting only to Domains Related to the Document's Topic

Efforts to minimize the number of possible authors arise from the fact that current algorithms do not achieve a high success rate for difficult problems which include:

- High number of possible authors: Hundreds to thousands authors make detection unreliable. A baseline is defined as $\frac{1}{|authors|}$, therefore, improving the baseline significantly is not equal to achieving a high accuracy.
- Comparing different styles of documents: Letters, discussion posts, blogs and articles use different key words (addressing, signature), stylistics (formal, informal), etc. Mixing such documents has negative impact on algorithms using word and n-gram frequencies.
- Comparing documents of different topics: For example, difference between political articles and personal correspondence is significant, therefore two political articles from different authors can be more similar to each other than an e-mail and a political article, both written by one author.
- Documents do not contain enough text: This particularly applies to discussions and e-mails which can contain only several sentences.

The disadvantage of filtering domains out is that the unknown document's author may not be among the authors intended to compare. But the unknown document can never be compared to all authors – despite the fact that people publish in the Internet under their real identities (advertisements, school assignments, social networks), it is impossible to access all these data (to enable web domains' parsing requires automatic approach and the total number of authors in the database would still be too high).

4 Authorship Identification Algorithms

This section describes the text preprocessing in the *ART* framework. The language of documents is detected and the text is tokenized, morphologically and syntactically annotated and disambiguated. These processes are language dependent, therefore, modules for each language are needed.

The main goal of the proposed framework is to support collecting data from Internet and to solve technical difficulties typical for the online environment, hence there is no restriction on the authorship identification algorithms.

Since the time of Mosteller and Wallace [8], a substantial amount of new research was done in the topic of authorship identification taking advantage of new findings in areas such as machine learning, information retrieval, and natural language processing. Current studies recommend to use machine learning methods that are effective for large data. The new contribution of the *ART* framework is that the process of collecting data can be accelerated and machine learning methods can work with more data to achieve better accuracy.

5 Retrieving Documents from the Internet

An intelligent exploitation of the documents retrieved from Internet needs a description of the format of the stored document meta-data. Since it is tedious to manually extract the structure of web pages to be able to download information about documents and authors, we propose a new, semi-automatic approach. It consists of the following 4 steps:

1. Firstly, a domain from which documents are going to be extracted is selected and visited by an operator. The operator manually registers to the domain using data describing her institution. Then it is necessary to submit a small number of documents $d_1, \ldots, d_k$ (e.g. discussion posts, blogs) while logged as the registered user.
2. In the second step, a crawler[1] is used to download web pages $P_1, P_2, \ldots$ in the domain until all pages containing information about documents $d_1, \ldots, d_k$ are found: $P_{d_1}, \ldots, P_{d_k}$.
3. In the third step, the HTML tree structure of the selected pages is detected by a HTML parser. Then minimal sequences of HTML tags are extracted to describe each attribute of the documents $d_1, \ldots, d_k$ using local search heuristics. It is important that each sequence describes the same information in all downloaded web pages, e.g. the title sequence defines a path to information about the title for every document. The structure of the domain is stored into the database as generated sequences of HTML tags.
4. Finally, all documents from the domain are downloaded by the crawler and processed. With the knowledge of the web page's structure, only data relevant for the authorship identification are collected and saved in the database. The algorithm is summarized in Table 2.

[1] a tool for web page analysis and download including the page links

In any future attempt to retrieve documents from already processed domain, $P_{d_1}, \ldots, P_{d_k}$ are downloaded again and their content is compared to the saved data in the database $(d_1, \ldots, d_k)$. If the content differs, either documents were edited (which is unlikely because documents were created by the operator), or the structure of pages in the domain was changed. Therefore, in this case all 4 steps are executed again. Otherwise, only new pages are processed.

Table 2. Domain structure identification

action	example

`Select domain`
`and register as author`
`and submit article`
	D = *www.domain.com*
	(Name, E-mail) $\rightarrow D$
	(Title, Text, Name) $\rightarrow D$

`Download domain texts`	*documents* $t_1, \ldots, t_n \in D$
`Search inserted document`	t_k = *(Title, E-mail, Name)*
`Extract structure of document`	$Title_k$: *body/div[@content]/h3* $Text_k$: *body/div[@content]/p* $Author_k$: *head/title*
`Process downloaded documents`	$Title_k$: *Introduction post* $Text_k$: *Text about web page's topic...* $Author_k$: *NLP Center*

6 Document Clustering According to the Authorship

Clustering of anonymous documents according to the authorship is very difficult. There even do not exist any recommended metrics for measuring the quality of a particular clustering in the authorship identification problem. We conducted some experiments but the results' accuracy was low. Although similarity of two documents can be compared with relatively high accuracy, for creation of large clusters many comparisons are made and even marginal errors decrease total accuracy significantly.

Therefore anonymous documents are not clustered and only documents signed by authors are put together. In order to adapt to data from an online environment for which identification of authors are not unique, an operation merging two clusters is allowed. It is very important to process data from different domains because the author's accounts may vary. Either it is a cosmetic change of identity (e.g. size of letters, leave out one word), or the author uses a completely different pseudonym.

Whenever a new author is inserted into the database, he or she is compared to each known author. If two authors' documents differ only marginally, their identities are connected. On the contrary, authors with same identities from different web domains are not linked automatically, their similarity has to exceed a specified limit that is more tolerant than in the case of two different identities. This is necessary because many pseudonyms and names overlap.

Despite the fact that the operation is time consuming, it is affordable because it is sufficient to apply it only to authors of documents with a similar theme and each author is processed only once.

7 Experimental results

Six evaluations were conducted to test the hypothesis that accuracy is improved by filtering out improbable authors. *ART* framework was used to automatically detect the structure of an extremist website *WM*[2] and to download all discussion posts (messages are mostly long 1 to 5 sentences).

Experiments are divided into two parts: In the part A, only documents of authors who wrote at least four documents are selected as test documents. At minimum three documents are used to create authors characteristics to which unknown document is compared. In the part B, author characteristic are generated from an arbitrary number of documents, therefore, the authorship recognition problem B is more difficult. Each evaluation is examined in Table 3 for part A and in Table 4 for part B. Each part consists of 3 scenarios:

- In the first case, documents (messages) were selected according to their extremist and racism theme. In the part A, authors who wrote less than three comments were filtered out because it is difficult to extract author's characteristics from very short texts. This scenario achieved the highest accuracy (22% and 6% documents were assigned to their author correctly in parts A and B).
- In the second scenario, data used in the first case were extended by small number of documents extracted from another website *I*[3]. Test documents remained same. The accuracy was lowered significantly to 3.1 and 0.7%.
- In the last scenario, data were further extended by big number of documents extracted from the website *I*. The accuracy was still 3.1 and 0.7%.

Results indicate that adding documents of different topic can substantially decrease the accuracy. If comparisons are not made under the same conditions (documents have different lengths, key words, levels of formality, . . .), authorship recognition algorithms perform worse. Furthermore, despite the fact that increasing number of authors' documents by adding messages of the same theme decrease the accuracy to a lesser extent than adding messages of different topic the performance decrease is still significant.

8 Conclusions

The *ART* framework extends previous works for authorship identification. The process of determining authorship is unchanged, but the manual documents obtaining is

[2] An extremist website `http://www.white-media.info/` was selected because it was mentioned recently in newspapers.

[3] Authors of discussion posts from the news server `http://idnes.cz` are added to the group of possible authors.

Table 3. Filtering domain experiment

Scenario 1A: Authors wrote at least 3 documents about the selected theme			
	Filtered topic	Other Topic	Total
Number of learn documents	144	–	144
Number of potential authors	8	–	8
Number of test documents	32	–	32
Accuracy	21.9%		

Scenario 2A: All documents are about the selected theme			
	Filtered topic	Other Topic	Total
Number of learn documents	189	98	287
Number of potential authors	8	58	66
Number of test documents	32	–	32
Accuracy	3.1%		

Scenario 3A: All documents are about the selected theme			
	Filtered topic	Other Topic	Total
Number of learn documents	189	946	1135
Number of potential authors	8	228	236
Number of test documents	32	–	32
Accuracy	3.1%		

replaced by more effective intelligent semi-automatic data retrieval, that has a positive impact on the quality of machine learning models used to detect an authorship. The proposed framework improves previous systems, because documents can be downloaded and processed from the web domain and regular updates are supported, therefore the authors' database is actual. In case of change of a web domain structure, the system is able to adapt to the change and retrieve documents without manual intervention.

Acknowledgments This work has been partly supported by the Ministry of the Interior of CR within the project VF20102014003 and by the Ministry of Education of CR within the Center of basic research LC536

References

1. A. Abbasi and H. Chen. Applying authorship analysis to extremist-group web forum messages. *Intelligent Systems, IEEE*, 20(5):67–75, 2005.
2. Ahmed Abbasi and Hsinchun Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005.
3. John Burrows. Delta': a measure of stylistic authorship 1. *Literary and Linguistic Computing*, 17(3):267–287, 2002.
4. Chaski,C. E. Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence*, 4(1):1–13, 2005.

Table 4. Filtering domain experiment

Scenario 1B: All documents are about the selected theme			
	Filtered topic	Other Topic	Total
Number of learn documents	189	–	189
Number of potential authors	37	–	37
Number of test documents	134	–	134
Accuracy	6%		

Scenario 2B: Small number of documents from another domain is added			
	Filtered topic	Other Topic	Total
Number of learn documents	189	98	287
Number of potential authors	37	58	95
Number of test documents	134	–	134
Accuracy	0.71%		

Scenario 3B: Many documents from another domain are added			
	Filtered topic	Other Topic	Total
Number of learn documents	189	946	1135
Number of potential authors	37	228	365
Number of test documents	134	–	134
Accuracy	0.7%		

5. Hsinchun Chen. Exploring extremism and terrorism on the web: the dark web project. In *Proceedings of the 2007 Pacific Asia conference on Intelligence and security informatics*, PAISI'07, pages 1–20, Berlin, Heidelberg, 2007. Springer-Verlag.
6. Jakubíček, Miloš - Horák, Aleš. Punctuation Detection with Full Syntactic Parsing. *Research in Computing Science, Special issue: Natural Language Processing and its Applications*, 46:335–343, 2010.
7. Koppel, Moshe - Schler, Jonathan - Argamon, Shlomo. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.*, 60:9–26, January 2009.
8. F. Mosteller and D. L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1964.
9. E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.
10. Sresha Yadav and Smita Jha. A framework for authorship identification of questioned documents: Forensic and linguistic convergence by. *MJAL*, 3(1):1–7, 2001.
11. Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, 2006.
12. Rong Zheng, Yi Qin, Zan Huang, and Hsinchun Chen. Authorship analysis in cybercrime investigation. In *Proceedings of the 1st NSF/NIJ conference on Intelligence and security informatics*, ISI'03, pages 59–73, Berlin, Heidelberg, 2003. Springer-Verlag.