Report from the PAN'14 Evaluation Lab at CLEF'14

Benno Stein and Martin Potthast

Bauhaus-Universität Weimar
99432 Weimar, Germany
`www.webis.de`

Francisco Rangel

Autoritas Consulting
Lorenzo Solano Tendero 7, 28043 Madrid, Spain
`http://www.autoritas.net/en`

Paolo Rosso

Universidad Politécnica de Valencia
Camino de Vera s/n. 46022 Valencia, Spain
`http://users.dsic.upv.es/~prosso/`

Efstathios Stamatatos

University of the Aegean
83200 - Karlovassi, Samos, Greece
`http://www.icsd.aegean.gr/lecturers/stamatatos`

Walter Daelemans and Ben Verhoeven

University of Anwerp
Prinsstraat 13, B-2000, Antwerpen, Belgium
`http://www.clips.ua.ac.be`

`http://pan.webis.de`   `pan@webis.de`

Abstract

We report on the outcomes of the PAN evaluation lab on uncovering plagiarism, authorship, and social software misuse, which was held in conjunction with the CLEF'14 conference. A total of 44 researchers attended the workshop and two keynote speakers were invited to talk about plagiarism detection and personality detection. In a steering committee meeting, the organizers discussed future directions of the lab, as well as the insights gained from intensive discussions with participants.

Summary

The PAN workshop is well-known for its contributions to the field of digital text forensics. In particular, our current focus is on plagiarism detection, author identification, author profiling, and social software misuse. The workshop brings together experts from science and industry, who discuss and compare their most recent results on the aforementioned fields. Since 2007, when PAN was first hosted at the SIGIR conference, it has attracted contributions from hundreds of researchers worldwide. In 2009 we have started to organize shared tasks at PAN, where we invite participants to work on selected problems of interest for which we hand out tailored test data sets. Participants then carry out research to solve these problems automatically and submit the results obtained for evaluation. This way, PAN has become one of the top venues for research on the aforementioned fields as well as for its comparative evaluation.

The 2014 edition of PAN featured four tasks, namely source retrieval, text alignment, authorship verification, and age/gender prediction. These tasks were carried out throughout the year, whereas in September 2014 we organized our annual workshop at the CLEF conference, where participants were given a chance to present their results. A total of 58 teams submitted their softwares to PAN for evaluation, and almost all of them also submitted a notebook paper describing their approach. A total of 44 people registered for the workshop, representing most of the participating teams: we invited all participants to present their work in the form of a poster presentation, and the authors of the best-performing softwares to also give talks about their approach. PAN was organized within three two-hour sessions at the conference. We dedicated some of this time to two keynote speakers, Paul Clough and Fabio Celli, who were invited to give a broader perspective and to share insights from their respective areas of expertise. Tables 1-3 overview the workshop program. In what follows, we briefly review the aforementioned activities and the obtained results.

Shared Tasks on Digital Text Forensics at PAN

The four shared tasks at PAN attracted researchers and practitioners from all over the world. For each of the tasks, the respective task organizing committee wrote an overview paper that goes into details about what participants did and what insights have been gained; we provide a brief summary of each task:

- *Source Retrieval.* In plagiarism detection, source retrieval is about searching for likely sources of a suspicious document. This task was organized for the third time at PAN, and it is one the tasks with the most advanced evaluation setup. For example, we have developed and maintain a large-scale web search engine that indexes a static web crawl so that participants need not worry about this step of the pipeline, and so that results can be reproduced even after years. Six softwares have been submitted and evaluated for this task. Altogether, it was found that current approaches are a little too focused on optimizing precision compared to recall. Also runtime should probably not be the key metric to optimize. A reasonable assumption probably is that recall is most important to the end user of a source retrieval system [1]. This was also the consensus of the discussions at the workshop about this task.

- *Text Alignment.* Another important subtask of plagiarism detection is text alignment, which is about matching passages of reused text between pairs of documents. Eleven teams submitted softwares for this task. The main challenge with this task for participants is to make their approaches robust against the various ways in which reused passages of text may be concealed by humans. The results of the evaluation indicate that the most difficult ways to do so include paraphrasing and summarization which each require tailored

2

approaches. This insight, in turn, prompted participants to develop methods to combine variants of their algorithms to work together simultaneously in a hybrid form, returning only the results obtained from that strategy which seems to perform best on a case by base basis [1]. At the workshop, also the new performance measures which have been developed for this task and employed for the first time have been a point of discussion. It remains an open question how to judge the performance of a performance measure, when the only reference point one has are existing performance measures.

- *Authorship Verification.* Within author identification, authorship verification is a task that aims at answering the question whether two given documents have the same author or no. The test data that was collected and compiled for this task has been significantly larger than in previous years. Moreover, data has been collected both for 4 different languages and 4 different genres, making the test data a lot more diverse [3]. This was well-received by participants; 13 softwares have been submitted to this task. At the workshop, discussions were mostly about how the author identification task should be continued. Given that this task is one of the most traditional tasks in the field of author identification, and given the number of participants that take part each year, it was generally agreed that the task should not be stopped. However, the diversity of problem variants that can be tackled in this context is huge, whereas despite our efforts to enlarge the the test data sets, they are still comparably small. It was agreed that more effort should be spent both by organizers and participants to collect even more data.

- *Author Profiling.* This task is concerned with predicting an author's demographics from her writing. For example, an author's style may reveal her age and gender. As one of the youngest tasks that are organized at PAN, author profiling is still in a phase of its life cycle where different things are tried out in order to determine the best ways for evaluation. Nevertheless, the task has attracted strong participation from a total of ten teams who submitted their softwares. After the criticism about the noise found in last year's data, the task organizers worked hard to collect new data of higher quality and to clean the existing data sets [2]. The attendees of the workshop discussed mostly about which kinds of features are best suited for this task, and how different text genres affect prediction performance. Interestingly, the best performance was achieved on Twitter data, where the text samples are comparably brittle and noisy in itself.

For each of the shared tasks organized at PAN, two participants were invited to give an oral presentation. All other participants were invited to do a poster presentation; in the corresponding poster session, not only the attendees of the PAN workshop, but also the attendees of the entire CLEF conference were present, so that PAN participants had a great opportunity to spread the word about their contributions to a broad audience.

Keynotes

We invited two keynotes, each followed by a brief discussion. The keynote speakers—Paul Clough from The University of Sheffield, UK, and Fabio Celli from the University of Trento, Italy—are well-known researchers in their domains of expertise and they shared their unique insights into the digital text forensics and related areas.

Paul Clough is well known for his research on text reuse, a field closely related to that of plagiarism detection. In his keynote he raised the question of how to prove the ownership of a piece of text in the digital society: he was asked to assist in a text attribution problem where text and images were illegitimately reused from the web pages of a small business. When the offender was approached about committing possible plagiarism their response was "prove it". Clough described how he approached the problem of proving ownership and the challenges it entailed.

Moreover, he shared his experiences gained from working on the EPSRC-funded Measuring Text Reuse (METER) project with the UK Press Association, along with the current text attribution literature, that informed his textual analysis. Finally, he also demonstrated how freely available resources can be used to tackle this kind of problem.

Fabio Celli talked about a new and important subtask of author profiling, namely personality recognition from text. This task consists of the automatic classification of an authors' personality traits from pieces of text she wrote. Celli explained how a classifier's predictions can be compared against gold standard labels, obtained by means of personality assessments like the Big5 personality test. Until recently, the extraction of personality types was limited to blogs and offline texts, while in recent years there is a strong interest in the scientific community about the extraction of personality from various sources, such as online social networks, speech and video. Current approaches to personality recognition are based on supervised learning, but this has several limitations, for example the cost of data annotation, the lack of domain adaptability and multilinguality. Celli presented an unsupervised method for personality recognition from text and some of its applications in social network analysis as well as in other NLP tasks.

Steering Committee Meeting

For the steering committee meeting, all of PAN's main organizers and most of the task organizers were present. All of the results obtained have been discussed, which particularly includes the participants' criticism about certain aspects and details of each task. For PAN 2015, it was agreed to invite Fabio Celli to organize a task on personality detection as part of the author profiling task. Moreover, it was agreed to continue the initiative of collecting participants' softwares instead of only their output. This was well-received both by participants as well as organizers as a way to improve the reproducibility of shared tasks in general. Finally, with regard to organization, further advertisement of our tasks to the NLP community is necessary to connect better to them, since most of PAN's tasks are NLP tasks.

Impact and Future of PAN

In recent years, we found that PAN has had a sizable impact on the digital text forensics. All of PAN's co-organizers frequently hear people talk about PAN on various occasions, such as conference visits, etc. More measurable is the impact of PAN's publications, where overview papers as well as notebook papers are cited often. The best sign of impact, however, is the fact that PAN repeatedly attracts many practitioners and researchers from all over the world to participate. All of these things show that our efforts in organizing shared tasks are paying off. This year's edition of PAN had good impact, too, since many of PAN's participants already announced to participate again next year; 13 have already signed up. Despite the aforementioned points of criticism, the overall feedback the organizers received was very positive.

In the future, we intend to make further contributions to the evaluation of each of our tasks. This may also include the introduction of new, or the reformulation of long-standing problems in order to better pinpoint the challenging problems associated with them. For example, as pointed out above, we will tackle the problem of personality detection in text next year. Another important goal is sustainability: evaluations at PAN shall be developed to a point at which they can be run automatically. To this end, we are continuing the developments of our experimentation platform TIRA, which was used to handle the submitted softwares of all of PAN's participants this year.

Lab Program and Speakers

The following three tables overview the lab program of PAN, which is an excerpt of the CLEF conference program.

Table 1: PAN lab program on September 15.

Time	Session
10:30-12:30	**Conference papers: Session 1**
(15 min. talk)	Supporting More-Like-This Information Needs: Finding Similar Web Content in Different Scenarios *Matthias Hagen and Christiane Glimm*
12:30-13:30	Lunch
13:30-15:00 (15 min. talk)	**Conference papers: Session 2** Discovering Similar Passages Within Large Text Documents *Demetrios Glinos*
15:00-15:30	Break
15:30-18:00 (15 min. talk)	**Conference papers: Session 2** Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling *Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein*
18:00-19:30	**Joint Poster Session + Welcome Reception**
	A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014 *Miguel A. Sanchez-Perez, Grigori Sidorov, Alexander Gelbukh*
	A Set-Based Approach to Plagiarism Detection - Notebook for PAN at CLEF 2012 *Robin Küppers, Stefan Conrad*
	Heterogeneous Queries for Synoptic and Phrasal Search *Simon Suchomel and Michal Brandejs*
	Using Intra-Profile Information for Author Profiling *A. Pastor López-Monroy, Manuel Montes-y-Gómez, Hugo Jair Escalante, and Luis Villaseñor-Pineda*
	A Simple Approach to Author Profiling in MapReduce *Suraj Maharjan, Prasha Shrestha, and Thamar Solorio*
	A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF) *Mahmoud Khonji and Youssef Iraqi*

Table 2: PAN lab program on September 16.

Time	Session
10:30-12:30	**Keynote & Plagiarism Detection**, Chair: Benno Stein
10:30-11:30	Proving Ownership: The Case of "Wag in a Bag" *Paul Clough*
11:30-11:50	Overview of the 6th International Competition on Plagiarism Detection *Martin Potthast, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, and Benno Stein*
11:50-12:10	A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014 *Miguel A. Sanchez-Perez, Grigori Sidorov, Alexander Gelbukh*
12:10-12:30	Heterogeneous Queries for Synoptic and Phrasal Search *Simon Suchomel and Michal Brandejs*
12:30-13:00	Lunch
13:30-14:30	**Keynote & Author Profiling**, Chair: Paolo Rosso
13:30-14:30	Unsupervised Personality Recognition from Text: Possible Applications *Fabio Celli*
14:30-14:50	Overview of the 2nd Author Profiling Task at PAN 2014 *Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, Walter Daelemans*
14:50-15:10	Using Intra-Profile Information for Author Profiling *A. Pastor López-Monroy, Manuel Montes-y-Gómez, Hugo Jair Escalante, and Luis Villaseñor-Pineda*
15:10-15:30	A Simple Approach to Author Profiling in MapReduce *Suraj Maharjan, Prasha Shrestha, and Thamar Solorio*
15:30-16:00	Break
16:00-18:00	**Author Identification**, Chair: Francisco Rangel
16:00-16:20	Overview of the Author Identification Task at PAN 2014 *Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A. Sanchez-Perez, and Alberto Barrón-Cedeño*
16:20-16:40	A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF) *Mahmoud Khonji and Youssef Iraqi*
16:40-17:00	Author Verification: Exploring a Large set of Parameters using a Genetic Algorithm *Erwan Moreau, Arun Jayapal, and Carl Vogel*
17:00-18:00	Discussion

Table 3: PAN lab program on September 16 (cont'd).

18:00-19:30	**Joint Poster Session**

Expanded N-Grams for Semantic Text Alignment
Samira Abnar, Mostafa Dehghani, Hamed Zamani, and Azadeh Shakery

Evaluating Robustness for 'IPCRESS': Surrey's Text Alignment for Plagiarism Detection
Lee Gillam and Scott Notley

Hashing and Merging Heuristics for Text Reuse Detection
Faisal Alvi, Mark Stevenson, and Paul Clough

A Hybrid Architecture for Plagiarism Detection
Demetrios Glinos

Developing High-Resolution Universal Multi-Type N-Gram Plagiarism Detector
Yurii Palkovskii and Alexei Belov

Plagiarism Alignment Detection by Merging Context Seeds
Philipp Gross and Pashutan Modaresi

A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014
Miguel A. Sanchez-Perez, Grigori Sidorov, Alexander Gelbukh

Machine Translation Evaluation Metric for Text Alignment
Prasha Shrestha, Suraj Maharjan, and Thamar Solorio

Heterogeneous Queries for Synoptic and Phrasal Search
Simon Suchomel and Michal Brandejs

VEBAV - A Simple, Scalable and Fast Authorship Verification Scheme
Oren Halvani and Martin Steinebach

A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF)
Mahmoud Khonji and Youssef Iraqi

A Single Author Style Representation for the Author Verification Task
Cristhian Mayor, Josue Gutierrez, Angel Toledo, Rodrigo Martinez, Paola Ledesma, Gibran Fuentes, and Ivan Meza

Author Verification: Exploring a Large set of Parameters using a Genetic Algorithm
Erwan Moreau, Arun Jayapal, and Carl Vogel

A Language Independent Author Verifier Using Fuzzy C-Means Clustering
Pashutan Modaresi and Philipp Gross

A Trinity of Trials: Surrey's 2014 Attempts at Author Verification
Anna Vartapetiance and Lee Gillam

Using Intra-Profile Information for Author Profiling
A. Pastor López-Monroy, Manuel Montes-y-Gómez, Hugo Jair Escalante, and Luis Villaseñor-Pineda

Age and Gender Identification in Social Media
James Marquardt, Golnoosh Farnadi, Gayathri Vasudevan, Marie-Francine Moens, Sergio Davalos, Ankur Teredesai, Martine De Cock

A Simple Approach to Author Profiling in MapReduce
Suraj Maharjan, Prasha Shrestha, and Thamar Solorio

DAEDALUS at PAN 2014: Guessing Tweet Author's Gender and Age
Julio Villena-Román and José Carlos González-Cristóbal

19:30	Social Dinner

References

[1] Martin Potthast, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, and Benno Stein. Overview of the 6th International Competition on Plagiarism Detection. In Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij, editors, *Working Notes Papers of the CLEF 2014 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2014. URL `http://www.clef-initiative.eu/publication/working-notes`.

[2] Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the Author Profiling Task at PAN 2014. In Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij, editors, *Working Notes Papers of the CLEF 2014 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2014. URL `http://www.clef-initiative.eu/publication/working-notes`.

[3] Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A. Sanchez-Perez, and Alberto Barrón-Cedeño. Overview of the Author Identification Task at PAN 2014. In Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij, editors, *Working Notes Papers of the CLEF 2014 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2014. URL `http://www.clef-initiative.eu/publication/working-notes`.