

PAN'09: 3rd Int. PAN Workshop - 1st Competition on Plagiarism Detection

Satellite workshop of 25th SEPLN Conference on

Uncovering Plagiarism, Authorship and Social Software Misuse

Donostia-San Sebastián, September 10, 2009

<http://www.webis.de/pan-09>

About the PAN Workshop:

The workshop shall bring together experts and researchers around the exciting and future-oriented topics of plagiarism detection, authorship identification, and the detection of social software misuse. The development of new solutions for these problems can benefit from the combination of existing technologies, and in this sense the workshop provides a platform that spans different views and approaches. The following list gives examples from the outlined fields for which contributions are welcome, but not restricted to:

Plagiarism detection:

- * plagiarism detection in general, in Web communities and social networks, and cross-language plagiarism
- * identifying near-duplicate and versioned documents of all kinds: text, software, image, music, video
- * technology for high-similarity retrieval such as fingerprinting and similarity hashing

Authorship identification:

- * models for authorship identification, authorship attribution, and writing style
- * NLP- and knowledge-based retrieval models to capture personal traits and sentiment
- * Web forensics, community fraud, and new Web infringements

Social Software Misuse Detection:

- * uncovering serial sharing and lobbying
- * monitoring vandalism, trolling, or stalking
- * trust, psychological and personality-based user studies, social aspects of Web misuse

Background:

Plagiarism analysis is a collective term for computer-based methods to identify a plagiarism offense. In connection with text documents we distinguish between corpus-based and intrinsic analysis: the former compares suspicious documents against a set of potential original documents, the latter identifies potentially plagiarized passages by analyzing the suspicious document with respect to changes in writing style.

Authorship identification divides into so-called attribution and verification problems. In the authorship attribution problem, one is given examples of the writing of a number of authors and is asked to determine which of them authored given anonymous texts. In the authorship verification problem, one is given examples of the writing of a single author and is asked to determine if given texts were or were not written by this author. As a categorization problem, verification is significantly more difficult than attribution. Authorship verification and intrinsic plagiarism analysis represent two sides of the same coin.

"Social Software Misuse" can nowadays be noticed on many social software based platforms. These platforms like Blogs, sharing sites for photos and videos, wikis and online forums are contributing up to one third of new Web content. "Social Software Misuse" is a collective term for anti-social behaviour in online communities; an example is the distribution of spam via the e-mail infrastructure. Interestingly, spam is one of the few misuses for which detection technology is developed at all, though various forms of misuse exist that threaten the different online communities. Our workshop shall close this gap and invites contributions concerned with all kinds of social software misuse.

 About the Competition on Plagiarism Detection:

The detection of plagiarism by hand is a laborious retrieval task, a task which can be aided or made automatic. The PAN competition on plagiarism detection shall foster the development of new solutions in this respect.

The competition divides into two tracks:

* External Plagiarism Analysis. Given a set of suspicious documents and a set of potential source documents the task is to find all passages within the suspicious documents which have been plagiarised from one or more of the source documents.

* Intrinsic Plagiarism Analysis. Given a set of suspicious documents the task is to detect paragraphs in the documents which have not been written by its main author. No source documents are given in this task.

A large corpus of artificial plagiarism containing cases which have been obfuscated and/or translated will be released for the competition. A development corpus, to be used in developing a detection software, will be released two months before the competition starts, a competition corpus will be used to evaluate and compare detection software. The former will contain fully annotated plagiarism cases, the latter will not.

The success of plagiarism detection software will be measured in terms of its precision, recall, and granularity.

 Important Dates:

open	Notification of interest for participation
21.03.2009	Release of the development corpus
21.05.2009	Release of the competition corpus
07.06.2009	Submission deadline for the competition
15.06.2009	Notification of competition results
01.07.2009	Submission deadline for the papers
15.07.2009	Notification of reviews
01.08.2009	Submission deadline for final version of the papers
10.09.2009 (afternoon)	PAN Workshop

 Workshop Organization:

Benno Stein	Bauhaus University Weimar, Germany
Paolo Rosso	Universidad Politécnica de Valencia, Spain
Efstathios Stamatatos	University of the Aegean, Greece
Moshe Koppel	Bar-Ilan University, Israel

 Competition Organization:

Bauhaus University Weimar:
Benno Stein, Martin Potthast, and Andreas Eiselt

Universidad Politécnica de Valencia:
Paolo Rosso and Alberto Barrón Cedeño

 Contact:

pan09@webis.de

Information about workshop and competition can be found at:
<http://www.webis.de/pan-09>