Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection

PAN'07

Benno Stein, Bauhaus University Weimar
Moshe Koppel, Bar-Ilan University, Israel
Efstathios Stamatatos , University of the Aegean, Greece

benno.stein@medien.uni-weimar.de, moishk@gmail.com, stamatatos@aegean.gr

About

Goal of the workshop was to bring together experts and prospective researchers around the exciting and future-oriented topic of plagiarism analysis, authorship identification, and high similarity search. This topic receives increasing attention, which results, among others, from the fact that information about nearly any subject can be found on the World Wide Web.

At first sight, plagiarism, authorship, and near-duplicates may pose very different challenges; however, they are closely related in several technical respects:

- *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. Authorship verification and intrinsic plagiarism analysis represent two sides of the same coin.

- *Near-duplicate detection* is mainly a problem of the World Wide Web: duplicate Web pages increase the index storage space of search engines, slow down result serving, and decrease the retrieval precision. Near-duplicate detection relates directly to plagiarism analysis: at the document level, near-duplicate detection and plagiarism analysis represent also two sides of the same coin. For a plagiarism analysis at the paragraph level, the same specialized document models (e. g. shingling, fingerprinting, hashing) can be applied, where a key problem is the selection of useful chunks from a document.

Contributions

¿From the submissions the program committee accepted 7 research papers for presenting original work and covering the workshop themes. The presentations were split into three sessions and the workshop was attended by about 30 people. An interactive form of communication was encouraged with continuous discussions in between and after the presentations.

The workshop was opened by Efstathios Stamatatos (University of the Aegean), who gave an invited talk on The Class Imbalance Problem in Author Identification. First, Efstathios described the state-of-the-art in author identification techniques focusing on instance-based vs. profile-based approaches as well as on text representation methods. Then, he presented solutions to deal with class imbalance, an important problem in author identification especially in the framework of forensic applications, for both instance-based and profile-based approaches.

Einat Amitay presented her joint work with Silvan Yogev and Elad Yom-Tov (IBM Research, Israel) about an interesting phenomenon observed on the Web, which they call "Serial Sharers": authors who publish excessively, in multiple forms, and under different identities. She focused on analyzing patterns in the contributions of a relatively small group of people who author too much web content (serial sharers). Based on a collection of Web pages authored by a couple of thousand different users and a compression-based classification algorithm she presented a quite interesting approach for detecting the identity of web authors.

The second session begun with a paper on Forensic Authorship Attribution for Small Texts by Ol'ga Feiguina and Graeme Hirst (University of Toronto). Graeme presented an approach to identifying the author of short texts (200-1,000 words) examining the significance of new syntactic information (bigrams of syntactic labels). The proposed method performs better in forensic texts rather than literary texts. However, part-of-speech and lexical features are still useful especially when combined with the proposed bigrams of syntactic labels.

The second paper of this session by George Mikros and Eleni Argyri (University of Athens) analyzed Topic Influence in Authorship Attribution. Eleni Argyri presented a study to explore topic-independent features for authorship attribution. A wide range of stylometric features were examined including vocabulary richness variables, lexical features, sentence level measures, word-length measures, and character level measures. Results on a Modern Greek corpus showed that the majority of the examined features are correlated with topic. By removing all these features the author identification performance remains practically unaffected.

The last paper of this session by Jussi Karlgren and Gunnar Eriksson (Swedish Institute of Computer Science) entitled Authors, Genre, and Linguistic Convention was presented by Jussi. He proposed features for quantifying the use of clauses and adverbials within sentences taking into account sequence patterns rather than averaging pointwise non-contextual observations. Another basic claim of their work is that intrinsic rather than extrinsic evaluation should be used to estimate the optimal features for both genre detection and authorship attribution. To this end, they propose the Kullback-Leibler divergence. The afternoon session was opened by the paper Adaption of String Matching Algorithms for Identification of Near-Duplicate Music Documents by Mattias Robine, Pierre Hanna, Pascal Ferraro, and Julien Allali (Universite de Bordeaux 1). Mattias presented a plagiarism detection approach for symbolic music documents. The proposed method mainly considers melody but takes also into account elements of music theory to detect musically important differences. Several experiments using known cases of music plagiarism indicate the

viability of the proposed approach.

Finally, a work about Intrinsic Plagiarism Analysis with Meta Learning by Benno Stein and Sven Meyer zu Eissen (Bauhaus University Weimar) was presented by Sven. He focused on the case of plagiarism detection without a reference collection. Essentially, this implies the quantification of writing style changes within a document. Using a test corpus they showed that the unmasking technology for authorship verification (developed by Koppel and Schler) also works for pretty short texts. In particular Benno and Sven use unmasking to obtain further evidence whether certain portions of a text, which were extracted with style heuristics, stem from another author. As a real-world case Sven exhibited the results of discriminating the plagiarized sections of a German habilitation thesis.

Open Questions

One of the key issues for both authorship attribution and plagiarism analysis is the lack of benchmark corpora to evaluate the proposed approaches. There is need for publicly-available large corpora covering several natural languages and providing control over topic, genre, period, etc. Thus, it is not yet feasible to conduct a wide range of comparative evaluation of different methods on the same data.

Another interesting open question is the optimal set of features for representing text in the framework of stylistic analysis. So far, the proposed features range from low-level measures such as character n-grams to lexical features and high-level measures such as syntactic analysis features or sequential features. Moreover, it is not yet clear how certain factors such as genre, authorship, and theme affect text representation features. On the other hand, compression-based approaches offer a parameter-free solution to this problem. However, they usually require significantly higher computational cost and their performance in many cases are poor.

Perhaps the most crucial factor is text-length. Although there are now robust methods able to deal with short texts (with less than 1,000 words), it is yet impossible to estimate a certain limit of text-length for appropriately quantifying the stylistic properties of the text. The definition of such a shortest text unit would allow the efficient segmentation of long texts into chunks so that to enable the detection of changes of writing style within a document.

Finally, another open question is the development of systems for plagiarism analysis, authorship attribution, or authorship verification that are able to explain their decisions. To illustrate this further, an author identification system used in the framework of a forensic application should be able to determine why a certain person is considered quite likely to be the author of a particular document the way a forensic science human expert would do it. To this end, the stylistic analysis systems should be able to match certain textual patterns with high-level stylistic characteristics.

Future

It turns out that the most important use cases for the uncovering of plagiarism and authorship will come from the World Wide Web: the WWW is the world's largest document repository, communication platform, and exhibition place. And, the Web provides individual access, public

address—along with anonymity at the same time. Most of the participating researchers at the workshop declared to address Web issues explicitly in their current and future work.

Especially with the Web we see another kind malicious user behavior and its problems coming into the research focus: the detection of individual content manipulation, e. g. in the form of vandalism and edit wars. Though this a sort of "adversarial" behavior this kind of phenomena has not been addressed by the adversarial IR community: Adversarial Information Retrieval on the Web (see the workshop on the SIGIR'06 among others) deals primarily with mass phenomena and "pollution" problems such as link-bombing, comment or blog spam, or search engine spam and optimization.

Acknowledgements

We would like to thank the SIGIR, the authors and presenters, and the renowned experts who served on the program committee for their contributions to make this high-level and exciting workshop possible!

References

- The workshop program can be found at
  http://www.uni-weimar.de/medien/webis/research/pan-07/program.html
- The workshop proceedings can be found at
  http://ceur-ws.org/Vol-276