Large-scale Author Verification
Temporal and Topical Influences

Michiel van Dam and Claudia Hauff
Web Information Systems
Delft University of Technology
the Netherlands

1. Introduction

**Author verification task:**
Has author \(X\) (with known reference texts \(K_1..K_i\)) written text \(T\)?

Limitations of publicly available corpora such as PAN [1]:
* very small (a few dozen test cases),
* few languages covered,
* texts \(K_i..K_j\) and \(T\) are long, and matched in time & genre

Goals:
* pipeline to create large-scale corpora automatically,
* investigate factors of time and topic on the accuracy of author verification

2. Research questions

**Topic hypothesis:**
Short and topically diverse reference documents make the verification problem more difficult.

Common **topic-based assumption**:
Two texts about similar topics are biased towards being recognized as from the same author.

**Temporal hypothesis** [3]:
Authors’ writing style changes over time. Texts written within a short period of time are more aligned than texts written at very different times.

3. Corpus

**Required:**
* many authors,
* many topics,
* extended period of time

Wikipedia Revision History

Wikipedia Talkpages

Comment Extraction
(2,500-10,000 chars, denoised)

Test Set Creation
(1 reference text per test case)

4. Results

**Topical similarity** could be verified (Different < Similar).

Common **topic-based assumption** could not be verified (similar achieves highest acc.).

**Temporal hypothesis** could be verified (Similar > Different).

Additional experiments in French, German, Spanish and Greek: the trends hold!

5. Future Work

* In-depth analysis across languages
* Comparison of different author verification algorithms

References

1. PAN: http://pan.webis.de/ (benchmark series starting in 2009, still ongoing)