Web information Retrieval Mini Project

SeenuArun AndiRajendran
Dhiraj Eadara

Project Motivation

- CbPD or Citation based plagiarism detection compares citation patterns
- Citation patterns of two suspected documents
- There are two types of citations
  - Implicit - e.g.: Herbert et al.
  - Explicit - e.g.: [1], [15], [3-13]
- Detecting explicit is simpler than implicit
- Implicit references are hard to detect due to language
- Resolving implicit references improve performance

Citation Based Plagiarism Detection

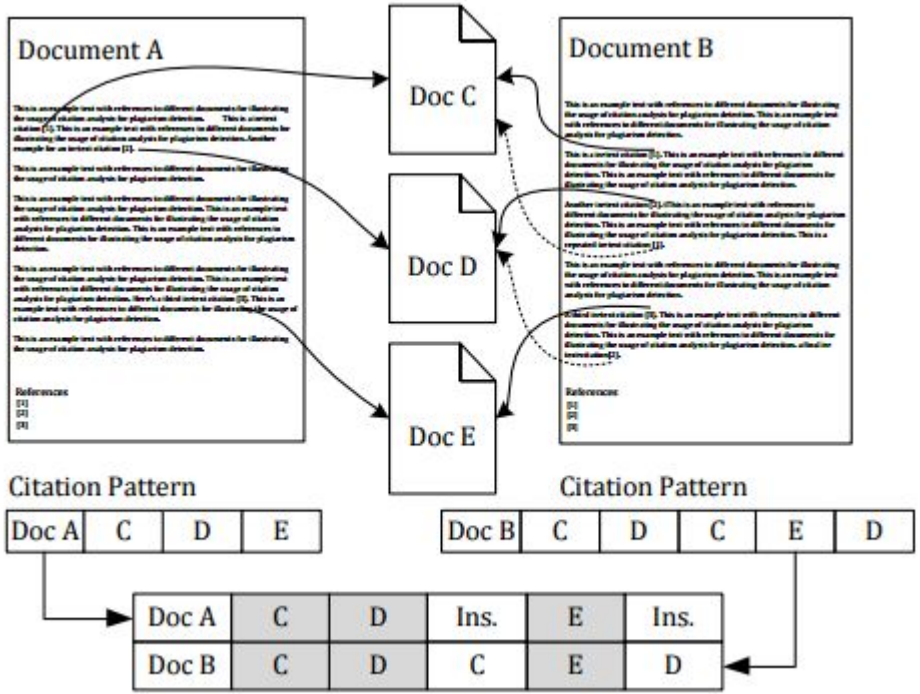

Figure credit:Gipp, Bela, and Norman Meuschke. "Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence." *Proceedings of the 11th ACM symposium on Document engineering*. ACM, 2011.

Our approach

- Identifies the bibliography of an academic document
- Tokenizes the bibliography into individual references
- Identifies author names within individual references
- Finds identified author names in the main text
- Applies pattern matching algorithms to detect pattern matches between different documents

Architecture

- Grobid - restfull service developed in java deployed by maven
- Post requests to the server
- Retrieving and parsing meta data results
- Searching text
- Retrieving implicit references in respective order

Results

- [**{'num': 2, 'AU': [u'Cancer', u'Center'], 'title': u'GermanCancerResearchCenter'}, {'num': 4, 'AU': [u'Herbert', u'Gerry', u'Mcqueen', u'Heid', u'Pfeufer', u'Illig', u'Wichmann', u'Meitinger', u'Hunter', u'Hu'], 'title': u'A common genetic variant is associated with adult and childhood obesity'},** {'num': 23, 'AU': [u'Walley', u'Asher', u'Froguel'], 'title': u'The genetic contribution to non-syndromic human obesity'}, {'num': 25, 'AU': [u'Heid', u'Huth', u'Loos', u'Kronenberg', u'Adamkova', u'Anand', u'Ardlie', u'Biebermann', u'Bjerregaard', u'Boeing'], 'title': u'Meta-analysis of the INSIG2 association with obesity including 74,345 individuals: does heterogeneity of estimates relate to study design? PLoS genetics'}]
- [**{'num': 4, 'AU': [u'Herbert', u'Gerry', u'Mcqueen', u'Heid', u'Pfeufer', u'Illig', u'Wichmann', u'Meitinger', u'Hunter', u'Hu'], 'title': u'A common genetic variant is associated with adult and childhood obesity'}, {'num': 9, 'AU': [u'Dina', u'Meyre', u'Samson', u'Tichet', u'Marre', u'Jouret', u'Charles', u'Balkau', u'Froguel'], 'title': u'Comment on "A common genetic variant is associated with adult and childhood obesity'}**, {'num': 14, 'AU': [u'Loos', u'Barroso', u'O 'rahilly', u'Wareham'], 'title': u'Comment on "A common genetic variant is associated with adult and childhood obesity'}, {'num': 16, 'AU': [u'Rosskopf', u'Bornhorst', u'Rimmbach', u'Schwahn', u'Kayser', u'Kruger', u'Tessmann', u'Geissler', u'Kroemer', u'Volzke'], 'title': u'Comment on "A common genetic variant is associated with adult and childhood obesity'}]

Challenges

1. Inconsistent or error-prone conversion of PDF to other formats
2. Wrong information stored amidst citation data.
3. Inconsistent retrieval of correct references
4. Difficulty in parsing the reference-related information for additional information such as author names, title, conference, etc.