Analyzing the information content of text-based files in supplementary materials of biomedical literature

Nona Naderi, Anaïs Mottaz, Douglas Teodoro, and Patrick Ruch
University of Applied Sciences and Arts of Western Switzerland (HES-SO)
Swiss Institute of Bioinformatics, Geneva, Switzerland
{firstname.lastname}@hesge.ch

We present an analysis of supplementary materials of PubMed Central (PMC) articles and show their importance in indexing and searching biomedical literature, in particular for the emerging genomic medicine field. On a subset of articles from PubMed Central, we use text mining methods to extract MeSH terms from abstracts, full-texts, and from text-based supplementary materials, such as spreadsheets and doc(x). We find that the recall of MeSH annotations increases about 5.9 percentage point (+20% on relative percentage) by considering supplementary materials compared to using only abstracts. We further compare the supplementary material annotations with annotations found in the articles' full-texts and we find out that the recall of MeSH terms increases by 1.5 percentage point (+3% on relative percentage). Additionally, we analyze genetic variant mentions in abstracts and full-texts and compare them with mentions found in text-based files in the supplementary materials. We find that the majority of variants (about 99%) are found in the text-based files of supplementary materials. Our study also highlights which types of information appear in text-based supplementary materials that are often missing in abstracts. In conclusion, we suggest that supplementary data should receive more attention from the information retrieval community, in particular in life and health sciences.