Text mining for HPC

Bruna Piereck Moura[1], Adriano Barbosa da Silva[2], Ana Maria Benko-Iseppon[1] e Ana Christina Brasileiro Vidal[1]

[1]Universidade Federal de Pernambuco - PPGG, [2]Luxembourg University, Luxembourg - LCSB

Each day, the volume of published data in biomedical and biological research is exponentially increasing, becoming a challenge to keep up dated the knowledge about a given topic of study. PubMed/Medline had approximately 22.7 mi citations until 2014/2015 and it has accomplished around to 26 mi citations until September, 2016. PESCADOR (Platform for Exploration of Significant Concepts Associated to co-Ocurrence Relationship) is an online flexible text mining tool that aggregate other programs to identify molecules pair of interactions. The objective of this work was to improve PESCADOR, turning possible HPC usage to analysing more data in less time. To achieve that, the tools and processes used by PESCADOR were individually adapted to HPC environment. At first, the 779 xml files from Medline database composed by approximately 30.000 citations each were automatically parsed to recover the PMID, Title and Abstract using python script. Followed by NLProt tagging tool, to highlight the protein and DNA names on xml parsing output (txt file). Analysing 10 files at time on one node (12 CPUs, 2 cores each) using the *parallel* tool, with 12 nodes-job, letting 2 free CPUs on each node. At last, based on MySQL and written in PHP language the LAITOR program, responsible for de interaction and co-occurrences identification, was modified, using python script, to SQLite format becoming able to run on HPC environment, all the libraries and index as much the query PHP lines at the original program were up date. LAITOR was run on one node, 8 files at time using *parallel* tool, letting 4 free CPUs. The xml parsing was finished after around 1h and was followed by the NLProt tagging, that was running for 5 days. At Last the LAITOR was running for around 3-4 days. The results of more than 23 mi abstracts were complete to be statistically analysed and curated. The same would not be possible in the online version of PESCADOR. All the adaptation needed took six (6) months. This results statistics and curation will make possible to enrich MESH terms, evaluate the well and poorly described protein interactions and estimate the needed time to curate all the interactions described until now.

Financial support: CAPES, CNPq, FACEPE