Abstract

As the enormous growth of information challenges the existing string analysis techniques for processing huge volume of data, there always seem to be a hope for newer inventions. Moreover, the problems encountered with the traditional methods such as low pruning power, increased false positives and poor scalability should be addressed with the appropriate solutions that cater to the need for improvement. Hence, this paper aims at proposing an improved similarity joins using SSPS MapReduce Framework that consists of a novel PSS Stemming algorithm along with three newly proposed filtering techniques such as SSize, SPositional and UI(Union –Intersection) that could effectively process large scale data by concerning the limitations of the traditional filtering methods. The experimentation shows that the framework reduces the false positives and run time cost with increased scalability than the existing frameworks.

References

Index Terms

Computer Science Information Sciences
Keywords

similarity joins, Hadoop, MapReduce, filtering and Verification