ABSTRAK

PENGEMBANGAN *PROGRESSIVE MINING OF SEQUENTIAL PATTERNS* (PISA) DENGAN MULTI BATASAN DAN SUMBER DATA HETEROGEN PADA PRAPROSES UNTUK KLASIFIKASI

Oleh
Regina Yulia Yasmin
NIM: 33211302
(Program Studi Doktor Teknik Elektro dan Informatika)

Kebutuhan untuk mendapatkan pengetahuan dari data yang berasal dari sumber heterogen dengan volume yang semakin membesar memberikan tantangan tersendiri pada proses penambangan data. Untuk mempercepat proses utama pada penambangan data, yaitu klasifikasi, penelitian ini merepresentasikan data ke dalam bentuk pola sekuens pada tahapan praproses sebagai input dari klasifikasi. Metoda yang diusulkan untuk mendapatkan pola sekuens dinamakan dengan *multiple constraint-based progressive mining of sequential patterns,* PISA**.

Penelitian dilakukan untuk mendapatkan solusi dari permasalahan riset, yaitu (1) bagaimana meningkatkan kinerja *sequential pattern mining* dalam menemukan pola sekuens yang memenuhi multi batasan untuk data dari sumber heterogen dengan struktur heterogen yang progresif secara efisien, (2) sejauh mana *sequential pattern mining* pada praproses dapat meningkatkan kinerja proses klasifikasi.

Tujuan penelitian adalah untuk menjawab persoalan peningkatan kinerja kecepatan dengan tetap mempertahankan akurasi dan skalabilitas pada proses klasifikasi untuk data dari sumber heterogen dengan struktur heterogen yang progresif, yang terdiri dari data terstruktur dan data tidak terstruktur, melalui pemanfaatan *progressive sequential pattern mining* dengan multi batasan pada tahap praproses. Kinerja kecepatan dan akurasi klasifikasi diukur dengan cara membandingkan parameter tersebut pada klasifikasi berbasis pola sekuens yang dihasilkan dari *Progressive mIning of Sequential pAtterns* (PISA) tanpa batasan, batasan tunggal dan multi batasan. Kinerja skalabilitas diukur dengan membandingkan kecepatan proses pada jumlah data yang berbeda. Penelitian ini dibatasi pada data terstruktur dan data teks yang tidak terstruktur serta dimungkinkan adanya penambahan atau pengurangan data yang tidak digunakan lagi pada basis data orijinal.

Metoda yang diusulkan adalah *multiple constraint-based progressive mining of sequential patterns* atau PISA** dan klasifikasi berdasarkan pola sekuens yang dihasilkan dari PISA**. PISA** dikembangkan dari metoda *progressive mining*

of sequential patterns, PISA, yang merupakan metoda *progressive sequential pattern mining*. PISA** mampu melakukan pengecekan multi batasan, termasuk batasan berbasis waktu pada saat *traversing Progressive Sequential tree*, PS-tree yang tidak dapat dilakukan oleh PISA. Metoda ini juga mampu menghasilkan jumlah pola sekuens yang lebih sedikit dan karena tidak membangkitkan kandidat pola sekuens maka metoda ini dapat menjaga skalabilitas untuk penambahan data. Pola sekuens yang jumlahnya sedikit dan sesuai dengan kebutuhan pengguna dapat meningkatkan kinerja klasifikasi untuk data dari sumber heterogen dengan struktur heterogen yang progresif.

Klasifikasi berbasis pola sekuens dikembangkan dari metoda *Classify-By-Sequence*, CBS yang memanfaatkan pola sekuens hasil dari metoda *sequential pattern mining* AprioriLike. Usulan metoda klasifikasi dinamakan CBS_CLASS**. CBS_CLASS** menggunakan PISA** untuk mendapatkan pola sekuens dan menggunakan pola sekuens tersebut sebagai *Classifier Sequential Patterns*, CSP, yang berfungsi sebagai fitur klasifikasi. Oleh karena kemampuan PISA** dalam menghasilkan pola sekuens yang jumlahnya sedikit dan sesuai dengan kebutuhan pengguna, maka CBS_CLASS** memiliki kinerja waktu klasifikasi yang meningkat dan tetap mempertahankan akurasi dan skalabilitas dibandingkan dengan CBS_CLASS yang menggunakan AprioriLike maupun PISA.

Eksperimen yang dilakukan bertujuan untuk mengukur kinerja *sequential pattern mining* pada praproses dan dampaknya pada perbaikan kinerja proses klasifikasi. Eksperiman dilakukan dalam 2 tahap, yaitu (1) untuk membandingkan kinerja akumulasi waktu eksekusi dan jumlah pola sekuens antara algoritma PISA, PISA* (PISA dengan batasan tunggal) dan PISA**(PISA dengan multi batasan), dan (2) untuk membandingkan kinerja waktu klasifikasi dan akurasi klasifikasi antara algoritma CBS_CLASS, CBS_CLASS* dan CBS_CLASS**. Data uji menggunakan data transaksi penjualan dan data ulasan produk dari perusahaan *e-commerce*. Hasil penelitian PISA** pada praproses menunjukkan bahwa akumulasi waktu yang dibutuhkan untuk pengecekan multi batasan adalah maksimum waktu dari pengecekan batasan dan jumlah pola sekuens hasil dari PISA** lebih rendah daripada PISA. Hasil penelitian pada klasifikasi memperlihatkan bahwa pada CBS_CLASS**, waktu klasifikasi berkurang dan akurasi klasifikasi dapat dipertahankan yang dibandingkan dengan CBS_CLASS dan CBS_CLASS*. Pada CBS_CLASS**, waktu klasifikasi berkurang rata-rata sebesar 87% dan akurasi klasifikasi meningkat rata-rata sebesar 7% dibandingkan dengan CBS_CLASS pada *minimum support* di bawah 0,4. Hal ini membuktikan hipotesis bahwa metoda PISA** terbukti menurunkan jumlah pola sekuens dan menjaga skalabilitas dari sistem. Selain itu, metoda ini meningkatkan kinerja kecepatan dan akurasi dari proses klasifikasi pada CBS_CLASS**.

Kata Kunci: multi batasan, sumber heterogen, struktur heterogen, progresif, data kompleks, *sequential pattern mining, multiple constraint-based progressive mining of sequential patterns, PISA**, CBS_CLASS**.*

ABSTRACT

THE DEVELOPMENT OF MULTIPLE CONSTRAINT PROGRESSIVE *MINING* OF SEQUENTIAL PATTERNS (PISA) WITH HETEROGENEOUS DATA SOURCES IN PREPROCESSING FOR CLASSIFICATION

By

Regina Yulia Yasmin

NIM: 33211302

(Doctoral Program in Electrical Engineering and Informatics)

*The necessity to gain knowledge from data that come from various sources with increasingly large volumes provides challenge in data mining. To accelerate the classification process, this study represents data into sequential patterns in preprocessing as input for classification. The proposed method PISA**, is a multiple constraint-based progressive mining of sequential patterns method to obtain sequential patterns.*

The study was conducted to solve research problems, (1) how to improve the performance of sequential pattern mining in finding sequential patterns that satisfy multiple constraints for heterogeneous sources, heterogeneous structures and progressive data efficiently, (2) how far sequential pattern mining in preprocessing can improve the performance of classification process in data mining.

The research objective is to address the need of increased speed performance while maintaining the accuracy and scalability of the classification process for heterogeneous sources, heterogeneous structures and progressive data, which consist of structured and unstructured data, through the use of progressive sequential pattern mining with multiple constraints in preprocessing. Classification's speed and accuracy performance was measured by comparing these parameters between classification based on sequential patterns from Progressive mIning of Sequential pAtterns (PISA) without constraint, single constraint and multiple constraints. Scalability performance was measured by comparing the processing speed of different number of data. This study was limited to structured and unstructured text data and there can be data addition or deletion along the process.

*The proposed method is multiple constraint-based progressive mining of sequential patterns or PISA** and classification based on sequential patterns generated from PISA**. PISA** was developed from progressive mining of sequential patterns, PISA, which was an progressive sequential pattern mining method. PISA** is capable of checking multiple constraints, including time-based*

constraints while traversing Progressive Sequential tree, PS-tree which can not be done by PISA. The method is also capable of producing lesser number of sequential patterns and since it does not generate sequence candidates, this method can maintain its scalability to data addition. The less number of sequential patterns and its conformity to user needs can improve the classification performance for heterogeneous sources, heterogeneous structures and progressive data.

*Classification based on sequential patterns was developed from Classify-By-Sequence, CBS method that used sequential patterns from sequential pattern mining AprioriLike. The proposed classification method was called CBS_CLASS**. CBS_CLASS** uses PISA** to get sequential patterns and utilized them as Classifier Sequential Patterns, CSP, which act as classification features. Therefore PISA**'s ability to generate lesser number of sequential patterns that conform to user needs, make CBS_CLASS** increase the classification time performance while maintaining accuracy and scalability compared to CBS_CLASS that uses AprioriLike or PISA.*

Experiments were aimed to measure the performance of sequential pattern mining in preprocessing and its impact on classification performance improvements. Experiments were carried out in two stages, (1) to compare the performance of accumulated execution time and the number of sequential patterns between PISA, PISA (PISA based on single constraint) and PISA** (PISA based on multiple constraints), and (2) to compare the performance of classification time and accuracy between CBS_CLASS, CBS_CLASS* dan CBS_CLASS**. Test data consisted of the sales transaction and product reviews from e-commerce company. The experiment results using PISA** in preprocessing showed that the accumulation execution time required for checking multiple constraints was the maximum time needed for multiple constraints checking and the number of sequential patterns from PISA** were lower than PISA. The experiment results on classification also showed that using CBS_CLASS**, the classification time was reduced and the classification accuracy was still maintained, compared to CBS_CLASS and CBS_CLASS*. On CBS_CLASS**, the classification time was reduced about average of 87% and the classification accuracy was increased about average by 7% compared to CBS_CLASS under the minimum support 0.4. These results proves the hypothesis that the PISA** reduces the number of sequential patterns and maintain system's scalability. Moreover, it increases classification's speed and accuracy performance of CBS_CLASS**.*

*Keywords: multiple constraints, heterogeneous data sources, heterogeneous data structures, complex data, sequential pattern mining, multiple constraint-based progressive mining of sequential patterns, PISA**, CBS_CLASS**.*