Prosiding Seminar Nasional Riset Information Science (SENARIS)
ISSN: 2686-0260
September 2019
Hal:1155-1164

Clustering Time Series Data Mining dengan Jarak Kedekatan Manhattan City

Relita Buaton[1], Muhammad Zarlis[2], Herman Mawengkang[2], Syahril Efendi[2]
[1]Graduate Program Of Computer Science,
[2]Department Of Computer Science
Faculty of Compurer Science and Information technology,
Universitas Sumatera Utara, Medan, Indonesia
bbcbuaton@gmail.com

Abstract*- The development of information technology is very rapid and is supported by the development of storage media technology and its application to all fields that produce huge amounts of data stacks generated from various sources, therefore need new techniques in managing data stacks. Data mining has become very important as an object and research study at this time because there are many data stacks found in agencies. Data mining is an analytical process of knowledge discovery in large and complex data sets. In this study the technique used is to conduct time series data mining clusters, using proximity to manhattan city. The time series graph is carried out by the sliding window to produce an analysis of the window for each cluster result. Based on cluster results, an analysis of knowledge transformation is carried out into new knowledge obtained from data mining time series data.*

Keywords*: data mining cluster time series*

Abstrak- Perkembangan teknologi informasi yang sangat pesat dan didukung oleh perkembangan teknologi media storage serta penerapannya pada semua bidang yang menghasilkan tumpukan data dengan kapasitas yang sangat besar yang dihasilkan dari berbagai sumber, oleh sebab itu perlu teknik baru dalam mengelola tumpukan data. Data mining menjadi sangat penting sebagai objek dan kajian penelitian saat ini karena banyaknya ditemukan tumpukan data dalam instansi. Data mining adalah proses analitis tentang penemuan pengetahuan dalam kumpulan data dengan jumlah yang besar dan kompleks. Dalam penelitian ini teknik yang digunakan adalah melakukan cluster time series data mining, dengan menggunakan jarak kedekatan manhattan city. Pada grafik time series dilakukan sliding window untuk menghasilkan analisis window untuk masing-masing hasil cluster. Berdasarkan hasil cluster, dilakukan analisa transformasi pengetahuan menjadi pengetahuan baru yang diperoleh dari data time series data mining.

Kata kunci*:cluster time series data mining*

1. PENDAHULUAN

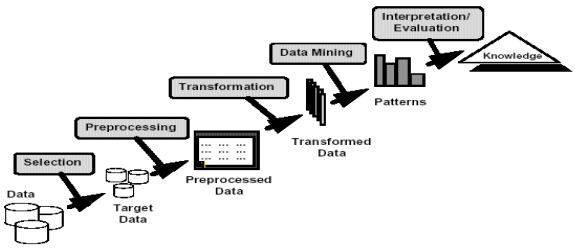
1.1. Data Mining

Data mining adalah proses analitis dari penemuan pengetahuan dalam kumpulan data yang besar dan kompleks, data mining adalah disiplin ilmu yang berada dalam interseksi statistik dan ilmu komputer. Lebih tepatnya, data mining adalah hasil hibridisasi statistik, ilmu komputer, kecerdasan buatan dan pembelajaran mesin [1]. Banyak ilmuwan data ingin mengeksplorasi data, mencari informasi untuk pengetahuan yang diperoleh melalui proses pengelompokan, klasifikasi, penemuan aturan, asosiasi dan visualisasi dalam penambangan data. Dalam statistik, Time Series adalah salah satu topik yang selalu dikaitkan dengan peramalan melalui serangkaian data yang tergantung pada periode waktu. Serangkaian berkala adalah kumpulan pengamatan yang dibuat secara kronologis. Data dari seri berkala memiliki karakteristik seperti besar, dimensi tinggi dan pembaruan terus menerus. Karakteristik berikutnya adalah bahwa sifat numerik dan kontinu dari data selalu dipandang sebagai keseluruhan daripada numerik

ISSN: 2686-0260
September 2019
Hal:1155-1164

individual. Oleh karena itu, tidak seperti database tradisional di mana pencarian kesamaan didasarkan pada pencocokan, pencarian kesamaan dalam data seri periodik didasarkan pada pendekatan. Contoh-contoh terkenal termasuk harga saham harian di Bursa Efek Jakarta, jumlah penggunaan ponsel setiap jam di Medan, dan pembacaan permukaan laut harian di Samudera Pasifik. Banyak penelitian telah dilakukan secara berkala berdasarkan kesamaan [2].

Menurut Kusrini dan Emha Taufiq Luthfi[3], *Data mining* adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam database. *Data Mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengektrasikan dan mengidentifikasikan informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Menurut Gartner Group *Data Mining* adalah suatu proses menemukan hubungan yang berarti, pola dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan tehnik pengenalan pola seperti tehnik statistik dan matematika. Data mining terdiri dari beberapa tahapan yaitu:



Gambar 1. Tahapan *Data Mining*

1. Tahap Pembersihan Data / *Selection*
   Yaitu dimana dilakukan proses pemilihan data yang akan digali (*field* yang dibutuhkan dalam proses *data mining*).
2. Tahap *Preprocessing / Data Warehouse*
   Mengeliminasi data yang tidak konsisten. Contohnya menghapus data yang kosong.
3. Tahap *Transformasi / Task-relevant Data*
   Proses pengubahan data menjadi bentuk lain, seperti jenis kelamin yang diganti menjadi 1 dan 0.
4. Tahap *Data Mining*
   Data yang telah diolah diawal siap untuk digali, sehingga dapat menghasilkan informasi yang baru.
5. Tahap Evalusi
   Setelah didapatkan hasil dari penambangan dan penggalian data maka hasil dari pengolahan data tersebut harus dievaluasi.[4]

1.2. Penelitian Terkait

Dalam seri penambangan data berkala, masalah mendasar adalah bagaimana menyajikan seri data berkala. Salah satu pendekatan yang umum adalah mengubah deret berkala menjadi domain lain sehingga dimensi yang dikurangi diikuti oleh

ISSN: 2686-0260
September 2019
Hal:1155-1164

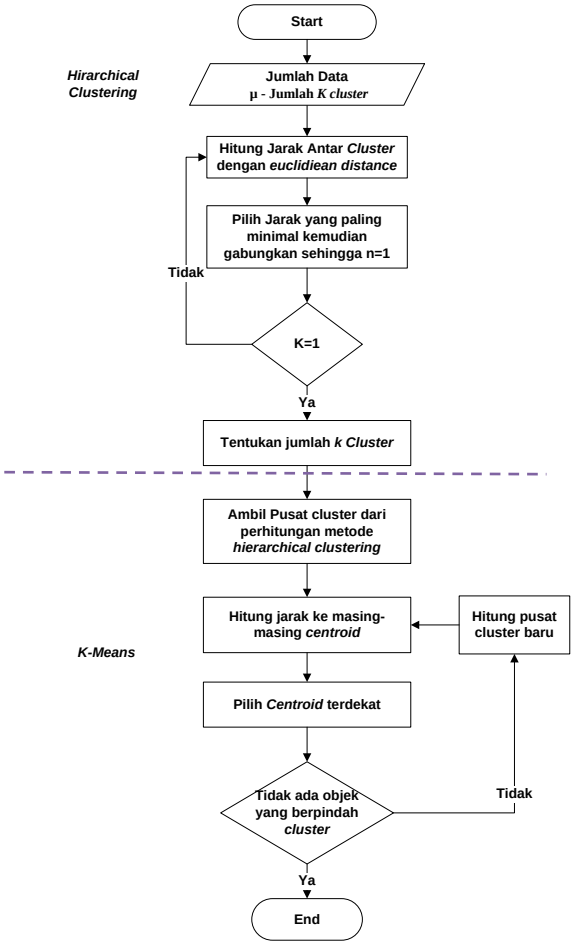
mekanisme indeks, penelitian deret waktu tidak optimal karena masih terbatas pada data tambang belum mampu mewakili deret waktu [5], mampu menemukan pola dalam data deret waktu[6], pola perlu dikembangkan untuk mengubah pola menjadi aturan. Aturan dapat ditemukan dari data deret waktu, tetapi masih dibatasi oleh overfitting. Selanjutnya, ukuran kesamaan dalam seri periodik atau sub sekuens dan proses segmentasi adalah dua tugas utama untuk berbagai tugas yang tercakup dalam penambangan sekuensial periodik. Salah satu tugas penambangan ini adalah aturan penemuan. [7]meneliti mekanisme menemukan aturan untuk seri periodik. Namun algoritma mereka hanya dievaluasi untuk kecepatan proses dan kemudian hanya pada data acak. Tidak diperlihatkan apakah algoritma ini menemukan aturan secara umum dalam seri periodik. [8] menggunakan representasi linear bagian demi bagian untuk mendukung aturan penemuan dalam seri periodik. Algoritma mereka diuji pada data keuangan, dengan prediksi yang tepat 68%. Metode yang paling banyak digunakan untuk penemuan aturan dalam literatur adalah dari [9]Mereka mengukur data dengan mengelompokkan K-means dari semua pelatihan dataset dan memasukkan data simbolik ke dalam algoritma asosiasi klasik dari penemuan aturan, kualitas aturan yang diinduksi dari data deret waktu dipengaruhi oleh parameter jumlah dan kluster. Keberhasilan suatu aturan diukur dengan menggunakan skor yang disebut ukuran-J. Tetapi pada telah ditunjukkan bahwa langkah kuantifikasi yang meliputi pengelompokan semua sub sekuens tidak akan dapat menghasilkan pusat kelompok. Masalah utama yang perlu ditangani dalam penambangan data berkala adalah jika hanya dengan visualisasi seri berkala yang dapat mencakup lebih dari ribuan pengamatan akan sangat sulit [10]. Bekerja dengan data mentah sangat tinggi akan sangat mahal dalam hal proses dan biaya penyimpanan. Oleh karena itu, kita memerlukan representasi atau abstraksi data tingkat tinggi. Penemuan aturan adalah salah satu cara melawan representasi. Sulit menyajikan data deret waktu dalam multi dimensi untuk ditambang

1.3. Clustering

Menurut [11], *Clustering* juga disebut sebagai *segmentation*. Metode ini digunakan untuk mengidentifikasi kelompok alami dari sebuah kasus yang di dasarkan pada sebuah kelompok atribut, mengelompokkan data yang memiliki kemiripan atribut. Selain itu *cluster analysis* merupakan pekerjaan mengelompokan data (objek) yang didasarkan hanya pada informasi yang ditemukan dalam data yang menggambarkan objek tersebut dan hubungan diantaranya. Tujuannya adalah agar objek – objek yang bergabung dalam sebuah kelompok merupakan objek – objek yang mirip (berhubungan) satu sama lain dan berbeda (tidak berhubungan) dengan objek dalam kelompok yang lain. Algoritma *K-Means* merupakan algoritma non hirarki yang berasal dari metode data *clustering*, Menurut [11]mengatakan bahwa metode *K-Means* ini mempartisi data kedalam kelompok sehingga data berkarakteristik sama dimasukan kedalam sat kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan kedalam kelompok yang lain. Adapun tujuan dari pengelompokan data ini adalah untuk meminimalkan fungsi objektif yang diset dalam proses pengelompokan, yang pada umumnya berusaha meminimalkan variasi didalam suatu kelompok dan memaksimalkan variasi antar kelompok.

ISSN: 2686-0260
September 2019
Hal:1155-1164

Dalam *hierarchical clustering* kita hitung jarak masing-masing obyek dengan setiap obyek yang lain. Selanjutnya kita temukan pasangan obyek yang jaraknya terdekat sehingga tiap obyek akan berpasangan dengan satu obyek atau dengan kelompok lain yang paling dekat jaraknya. Gambar 2 adalah flowchart yang menjelaskan urutan pengerjaan penelitian dengan menggunakan metode *hierarchical clustering* dan *Kmeans*.



Gambar 2. Algoritma *Hierarchical clustering* dan *K-means*

Pengelompokan data dengan metode *K-Means* ini secara umum dilakukan dengan cara sebagai berikut:

1. Tentukan jumlah kelompok,
2. Alokasikan data kedalam kelompok secara acak,
3. Hitung pusat kelompok (sentroid/rata- rata) dari data yang ada di masing – masing kelompok
4. Alokasikan masing – masing data ke centroid/rata-rata terdekat,
5. Kembali kelangkah 3, masih ada data yang berpindah kelompok, atau apabila ada perubahan nilai sentroid diatas nilai ambang yang ditentukan, atau apabila perubahan niai pada fungsi objektif yang digunakan masih diatas nilai ambang yang ditentukan.

Beberapa alternatif penerapan *K-Means* dengan beberapa pengembangan teori-teori penghitungan terkait telah diusulkan. Hal ini termasuk pemilihan:

ISSN: 2686-0260
September 2019
Hal:1155-1164

1. *Distance space* untuk menghitung jarak di antara suatu data dan *centroid*
   Beberapa *distance space* telah diimplementasikan dalam menghitung jarak (*distance*) antara data dan *centroid* untuk prinsip dasar rumus dalam perhitungan *istances* dan *Similarity Coeficients* untuk beberapa pasang dari item *Ecluidean Distance*:

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+...+(x_p-y_p)^2} \quad (1)$$

   Atau

$$d(x,y)=\left[\sum_{i=1}^{p}|\ x_i-y_i\ |^2\right]^{1/2} \quad (2)$$

   Sedangkan untuk *L2* (*Euclidean*) *distance space*, jarak antara dua titik dihitung menggunakan rumus sebagai berikut:

$$D_{L2}(x_2,x_1)=\left||x_2,x_1|\right|=\sqrt{\sum_{j=1}^{p}(x_{2j}-x_{1j})^2} \quad (3)$$

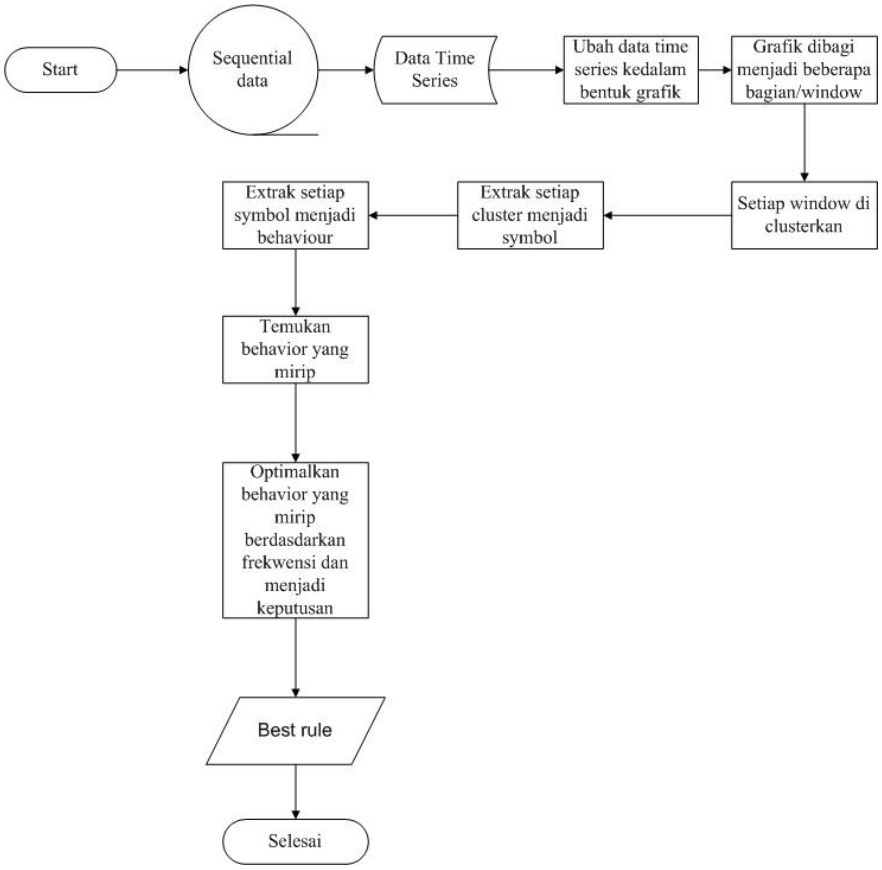
2. Metode pengalokasian data kembali ke dalam setiap *cluster*
   Secara mendasar, ada dua cara pengalokasian data kembali ke dalam masing-masing cluster pada saat proses iterasi clustering. Kedua cara tersebut adalah pengalokasian dengan cara tegas (hard), dimana data item secara tegas dinyatakan sebagai anggota cluster yang satu dan tidak menjadi anggota cluster lainnya, dan dengan cara fuzzy, dimana masing-masing data item diberikan nilai kemungkinan untuk bisa bergabung ke setiap cluster yang ada
3. *Objective function* yang digunakan.
   Objective function yang digunakan khususnya untuk *Hard K-Means* dan *Fuzzy K-Means* ditentukan berdasarkan pada pendekatan yang digunakan. Untuk metode *Hard K-Means*, *objective function* yang digunakan adalah sebagai berikut:

$$J(U,V))=\sum_{k=1}^{N}\sum_{i=1}^{c}a_{ik}D(x_k,v_i)^2 \quad (4)$$

2. METODE PENELITIAN

Untuk menghasilkan hasil cluster dengan tingkat similarity terbaik secara umum tahapan dan kerangka kerja penelitian yang digunakan adalah dengan mengembangkan kerangka penelitian yang telah dikembangkan oleh [12], secara umum memiliki 4 tahapan yakni proses transforasi data menjadi grafik time series 2 dimensi, melakukan pembagian grafik menjadi beberapa siliding window, pengelompokan data grafik, membuat pengetahuan dalam bentuk rule dari hasil analisa cluster grafik, yang ditampilkan pada diagram blok berikut ini

ISSN: 2686-0260
September 2019
Hal:1155-1164



Gambar 3. Kerangka Kerja Penelitian

3. HASIL DAN PEMBAHASAN

Proses optimasi yang digunakan dalam menghasilkan cluster time series data mining adalah analisa pengelompokan cluster pada sliding window time series. Data yang digunakan adalah time series temperatur suhu rata-rata. Dengan menggunakan model hasil yang telah ditemukan, berikut ditampilkan pembahasan sebagai berikut

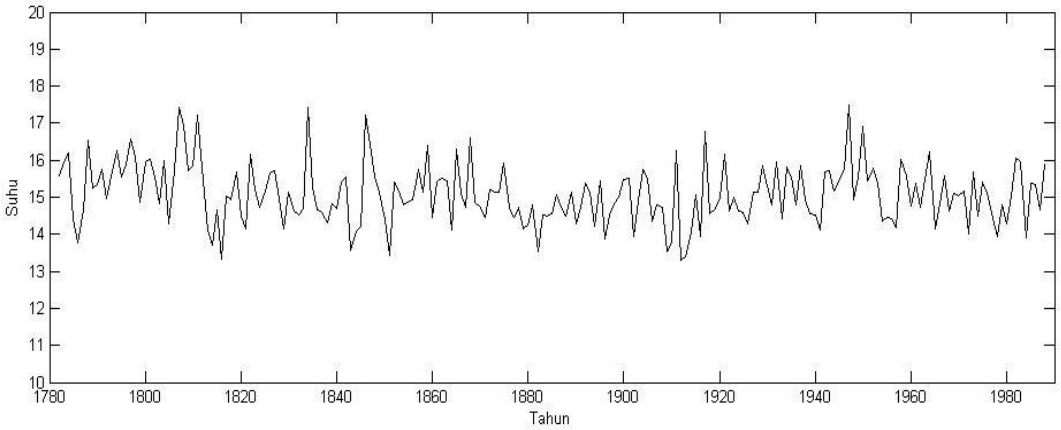
Tabel 1. Data Time Series Temperatur

Tahun	Suhu (^{0}C)	Tahun	Suhu (^{0}C)	Tahun	Suhu (^{0}C)	Tahun	Suhu (^{0}C)
1782	15,58	1885	14,57	1833	14,68	1936	14,8
1783	15,96	1886	15,06	1834	17,41	1937	15,84
1784	16,18	1887	14,72	1835	15,29	1938	14,93
1785	14,43	1888	14,5	1836	14,66	1939	14,54
1786	13,78	1889	15,13	1837	14,59	1940	14,51
1787	14,66	1890	14,3	1838	14,32	1941	14,1
1788	16,53	1891	14,72	1839	14,82	1942	15,68
1789	15,25	1892	15,37	1840	14,68	1943	15,72
1790	15,35	1893	15,13	1841	15,41	1944	15,18
1791	15,75	1894	14,23	1842	15,56	1945	15,45
1792	14,96	1895	15,46	1843	13,58	1946	15,76
1793	15,72	1896	13,89	1844	14,07	1947	17,49
1794	16,27	1897	14,52	1845	14,21	1948	14,94
1795	15,56	1898	14,83	1846	17,23	1949	15,54
1796	15,9	1899	15,01	1847	16,44	1950	16,9
1797	16,58	1900	15,47	1848	15,55	1951	15,44

ISSN: 2686-0260
September 2019
Hal:1155-1164

Tahun	Suhu (^{0}C)	Tahun	Suhu (^{0}C)	Tahun	Suhu (^{0}C)	Tahun	Suhu (^{0}C)
1798	16,09	1901	15,52	1849	15,12	1952	15,79
1799	14,88	1902	13,93	1850	14,37	1953	15,38
1800	15,97	1903	14,91	1851	13,44	1954	14,36
1801	16,02	1904	15,76	1852	15,42	1955	14,46
1802	15,55	1905	15,52	1853	15,16	1956	14,43
1803	14,84	1906	14,36	1854	14,81	1957	14,18
1804	15,99	1907	14,8	1855	14,88	1958	16,01
1805	14,29	1908	14,74	1856	14,93	1959	15,58
1806	15,72	1909	13,54	1857	15,74	1960	14,77
1807	17,43	1910	13,76	1858	15,14	1961	15,36
1808	16,94	1911	16,27	1859	16,39	1962	14,73
1809	15,73	1912	13,3	1860	14,46	1963	15,47
1810	15,87	1913	13,4	1861	15,4	1964	16,24
1811	17,21	1914	14,08	1862	15,5	1965	14,13
1812	15,43	1915	15,06	1863	15,45	1966	14,84
1813	14,1	1916	13,93	1864	14,1	1967	15,58
1814	13,72	1917	16,78	1865	16,3	1968	14,63
1815	14,65	1918	14,56	1866	15,09	1969	15,1
1816	13,34	1919	14,65	1867	14,72	1970	15,02
1817	15,04	1920	15	1868	16,62	1971	15,16
1818	14,93	1921	16,16	1869	14,87	1972	14
1819	15,68	1922	14,62	1870	14,72	1973	15,68
1820	14,5	1923	14,99	1871	14,44	1974	14,5
1821	14,13	1924	14,62	1872	15,22	1975	15,4
1822	16,17	1925	14,59	1873	15,14	1976	15,12
1823	15,22	1926	14,3	1874	15,14	1977	14,54
1824	14,71	1927	15,15	1875	15,91	1978	13,94
1825	15,16	1928	15,13	1876	14,71	1979	14,78
1826	15,65	1929	15,87	1877	14,46	1980	14,28
1827	15,71	1930	15,31	1878	14,71	1981	15,04
1828	14,98	1931	14,78	1879	14,13	1982	16,06
1829	14,15	1932	15,94	1880	14,24	1983	15,94
1830	15,12	1933	14,42	1881	14,78	1984	13,91
1831	14,64	1934	15,83	1882	13,53	1985	15,36
1832	14,51	1935	15,55	1883	14,53	1986	15,34
1884	14,48	1987	14,67	1988	15,88		

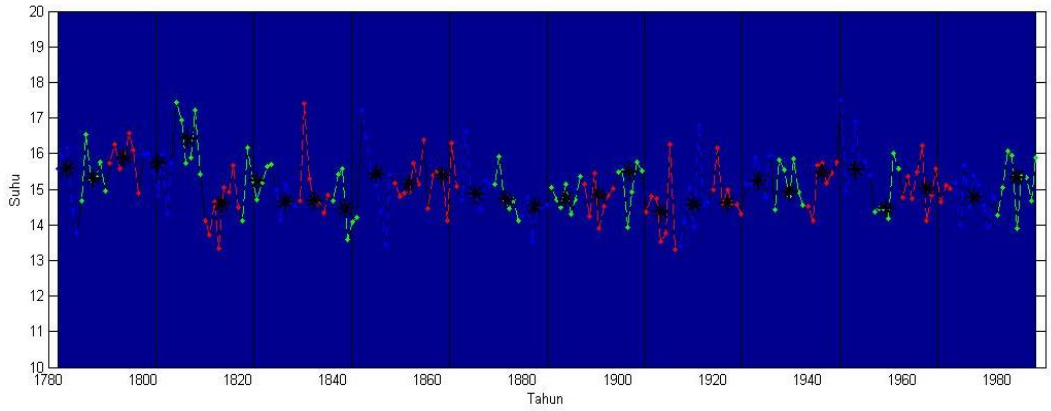
Data pada tabel 1 diproses dengan melakukan plot data menjadi grafik time series



Gambar 4. Grafik Time Series Temperatur

ISSN: 2686-0260
September 2019
Hal:1155-1164

Gambar 4. menunjukkan hasil plot data time series temperatur terhadap waktu(tahun), terlihat jelas bahwa grafiknya tidak linear terjadi perubahan bentuk seiring berjalannya waktu setiap tahun. Secara visual sangat sulit dianalisis jika menggunakan analisis pola, sulit mengidentifikasi aturan yang terkandung didalamnya dan potensial menarik. Langkah berikutnya membagi grafik menjadi beberapa window yang disebut dengan sub sequence time series, dalam kasus ini dibagi dalam 10 window dan selanjutnya window tersebut akan menjadi pusat analisis



Gambar 5. Hasil Analisis Window

Gambar 5 menunjukkan hasil analisis untuk setiap window, setiap window menghasilkan titik titik yang diperoleh melalui perhitungan jarak similarity, titik tersebut merupakan hasil trend untuk window seiring dengan terjadinya perubahan waktu, data setiap analisis window disajikan pada tabel 2

Tabel 2. Data Analisis Window

Window	**Tahun**	**Pusat Cluster**	**Pengetahuan Baru**
I	1780-1800	1784-15,6 ^{0}C	Suhu cenderung 15,3 ^{0}C sampai dengan 15,9 ^{0}C
		1796-15,9 ^{0}C	
		1789-15,3 ^{0}C	
II	1801-1820	1816-14,6 ^{0}C	Suhu cenderung 14,6 ^{0}C sampai dengan 16,4 ^{0}C, terjadi peningkatan suhu signifikan pada tahun 1801 yakni 16,4 ^{0}C
		1809-16,4 ^{0}C	
		1803-15,7 ^{0}C	
III	1821-1840	1836-14,7 ^{0}C	Suhu cenderung 14,6 ^{0}C sampai dengan 15,42 ^{0}C
		1824-15,2 ^{0}C	
		1830-14,6 ^{0}C	
IV	1841-1860	1856-15,1 ^{0}C	Suhu cenderung 14,4 ^{0}C sampai dengan 15,4 ^{0}C
		1842-14,4 ^{0}C	

ISSN: 2686-0260
September 2019
Hal:1155-1164

		1849-15,4 ^{0}C	
V	1861-1880	1863-15,1 ^{0}C	Suhu cenderung 14,7 ^{0}C sampai dengan 15,4 ^{0}C
		1876-14,4 ^{0}C	
		1870-15,4 ^{0}C	
VI	1881-1900	1896-14,8 ^{0}C	Suhu cenderung 14,5 ^{0}C sampai dengan 14,8 ^{0}C
		1889-14,7 ^{0}C	
		1882-14,5 ^{0}C	
VII	1901-1920	1990-14,4 ^{0}C	Suhu cenderung 14,4 ^{0}C sampai dengan 15,5 ^{0}C
		1902-15,5 ^{0}C	
		1916-14,6 ^{0}C	
VIII	1921-1940	1923-14,6 ^{0}C	Suhu cenderung 14,6 ^{0}C sampai dengan 15,2 ^{0}C
		1936-14,9 ^{0}C	
		1929-15,2 ^{0}C	
IX	1941-1960	1943-15,4 ^{0}C	Suhu cenderung 14,4 ^{0}C sampai dengan 15,5 ^{0}C
		1956-14,4 ^{0}C	
		1950-15,5 ^{0}C	
X	1961-1980	1965-15,0 ^{0}C	Suhu cenderung 14,8 ^{0}C sampai dengan 15,3 ^{0}C
		1984-15,3 ^{0}C	
		1975-14,8 ^{0}C	

4. KESIMPULAN

Berdasarkan analisa pengetahuan dari setiap window maka dapat ditentukan rule yang berpotensial menarik serta memangkas rule yang tidak menarik, rule yang cenderung memiliki redundansi, rule yang tidak memiliki redundansi, maka best rulenya adalah suhu cenderung pada angka 14 ^{0}C sampai dengan 15 ^{0}C setiap tahun, dan terjadi peningkatan suhu yang significan pada tahun 1809. Jarak kedekatan dengan menggunakan manhattan City mampu menemukan pengetahuan dengan analisis cluster.

DAFTAR PUSTAKA

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[2] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient Similarity Search In Sequence Databases," *Springer*, 1993.

ISSN: 2686-0260
September 2019
Hal:1155-1164

[3] K. Luthfi emha taufiq, *Algoritma Data Mining*. Yogyakarta: ANDI, 2009.
[4] R. Buaton, Y. Sundari, and Y. Maulita, "Clustering Tindak Kekerasan Keker Pada Anak Menggunakan Algoritm oritma K-Means Dengan Perbandingan ingan Jarak Kedekatan Manhattan City Dan an Eu Euclidean," vol. 1, no. 2, 2016.
[5] T. Fu, "Engineering Applications of Artificial Intelligence A review on time series data mining," *Eng. Appl. Artif. Intell.*, vol. 24, no. 1, pp. 164–181, 2011.
[6] E. Keogh, "Clustering of Time Series Subsequences is Meaningless : Implications for Previous and Future Research," 2005.
[7] S. Park and wesley Chu, "Discovery and matching elastic rules from sequence database," *Fundam. Inform.*, 2001.
[8] H. Wu, B. Salzberg, and D. Zhang, "Online Event-driven Subsequence Matching over Financial Data Streams," *Sigmod Conf.*, 2004.
[9] G. Das, K. Lin, and H. Mannila, "Rule discovery from time series," *Am. Assoc. Artif. Intell.*, 1998.
[10] J. Lin, S. Lonardi, J. Lin, and S. Lonardi, "Visualizing and Discovering Non-Trivial Patterns In Large Time Series Databases Short running title : Time Series Visualization Visualizing and Discovering Non-Trivial Patterns In Large Time Series Databases," 2005.
[11] E. Prasetyo, *Data Mining : Konsep Dan Aplikasi Menggunakan Matlab*. Yogyakarta: ANDI, 2012.
[12] K. U. Yoshiki Tanaka, Kazuhisa Iwamoto, "Discovery of Time-Series Motif from Multi-Dimensional Data Based on MDL Principle," *Springer Sci. + Bus. Media, Inc. Manuf. Netherlands*, no. 2000, pp. 269–300, 2005.