Developing Sequence Pattern of Diseases Using PrefixSpan Method Study Case: Dr. Soetomo General Hospital

Silvia Rostianingsih, Gregorius Satia Budhi, and Leonita Kumalasari Theresia

***Abstract*—Dr. Soetomo General Hospital has used computerized system to record their patient's diseases. With the large amount of data to be analyzed, Dr. Soetomo General Hospital needs to know the disease pattern to prevent and cure the disease. Based on the problem, the hospital needs to develop an application that generates sequences pattern of diseases so it could be used to predict sequence pattern of disease in later day.**

This application is built with PrefixSpan method to generate disease pattern in a particular region on particular time according Dr. Soetomo's General Hospital historical data. Output of this application is rules in the form of table and graph.

***Index Terms*—Data mining, disease, PrefixSpan, sequential pattern mining.**

I. INTRODUCTION

Dr. Soetomo General Hospital in Surabaya, Indonesia is a national hospital which acts as a reference from other hospitals. History of the patients is stored using *Oracle Data and Application* [1].

The increasing of civilization in East Java province is increased the patients with various type of disease. The hospital needs tool to monitor this occurrence in order to anticipate the spread of the disease.

This research is offering PrefixSpan sequential pattern mining to discover disease pattern from inpatient history with the disease's name and the occurrence region.

II. PREFIXSPAN METHOD

Sequential pattern mining is a method to discover the relation between items in a dataset [2].

Prefix-projected sequential pattern mining called PrefixSpan is a method to project sequence databases based on frequent prefixes because each subsequence frequent can be discover by growth frequent PrefikSpan. The PrefixSpan is using the following method [3]:

1) Scan S|α once, find the set of frequent items b such that:
   a. b can be assembled to the last element of α to form a sequential pattern; or
   b. <b> can be appended to α to form a sequential pattern
2) For each frequent item b, append it to α to form a sequential pattern α′, and output α′.
3) For each α′, construct α′-projected database S|α′, and call PrefikSpan (α′, l+1, S|α′).

This research is using bi-level projection calculation, which is:

1) Scan the sequence to get length-1 item.
2) Create triangular matriks from length-1 item.
3) For each length-2 sequential pattern, build a-projected database and count the occurrence item, then build s-matrix.
4) Each item is put in the end of length-2 sequential pattern [4].

III. DATA PREPARATION AND PROCESS

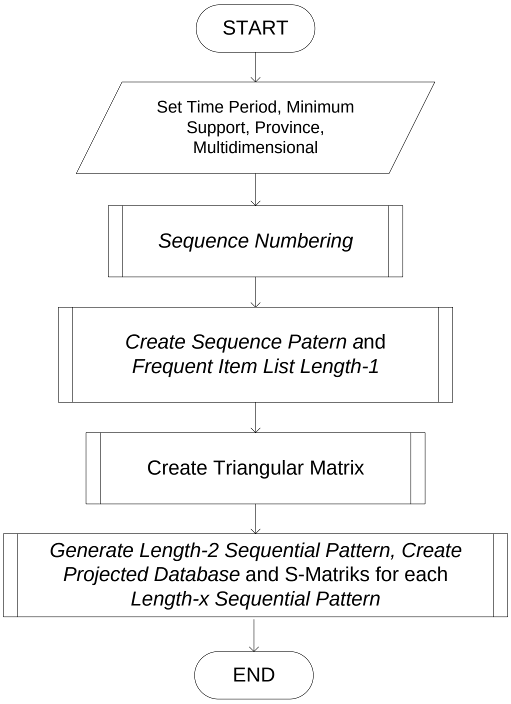

START
Set Time Period, Minimum Support, Province, Multidimensional
Sequence Numbering
Create Sequence Patern and Frequent Item List Length-1
Create Triangular Matrix
Generate Length-2 Sequential Pattern, Create Projected Database and S-Matriks for each Length-x Sequential Pattern
END

Fig. 1. PrefixSpan flowchart

This research is using patient table (regency, sex, date of birth, and other information), disease table (list of disease), diagnose table (time, doctor, patient, type of diagnose), province and regency table [5]. The method to generate rules (Fig. 1) consists of:

Manuscript received May 1, 2012;revised September 24, 2012. This work was supported in part by Directorate General of Indonesian Higher Education under Grant 110/SP2H/PP/DP2M/IV/2009, 326/SP2H/PP/DP2M/IV/ 2010, 0082/SP2H/PPK7/KL/IV/2011with title "Design and Development of Medical Record Data Warehouse Application System for Supporting RSU Dr. Soetomo Strategic Decisions.

Silvia Rostianingsih, Gregorius Satia Budhi and Leonita Kumalasari Theresia are with Petra Christian University, Siwalakerto 121-131 Surabaya, Indonesia (e-mail: silvia@peter.petra.ac.id).

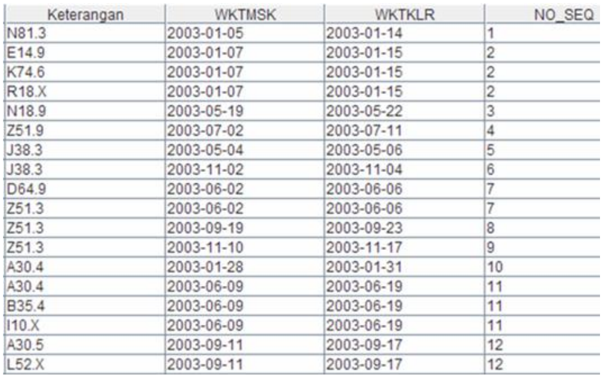
DOI: 10.7763/IJIMT.2012.V3.291

1) Sequence numbering which process based on id patient, time for in and out patient. Patient with the same id can have different sequence number because sequence number is based on period time of the last time patient out with the next time patient in.
2) Sequence Pattern is created from item in sequence. If an item is occurrence in recurrent, it is only written once. After the sequence pattern is created, the frequent of item (become frequent item list length-1) which fulfill the minimum support is counted. The result is a descending list from the frequent item.
3) Triangular Matrix can be created by built matrix with item length-1 number x item length-1 number size, which each cell is contain of three data which is present length-2 sequence pattern. Each data from cell which fulfills minimum support will be frequent item list length-2.
4) Each length-2 sequential pattern which fulfills minimum support will create projected database, length-1 in projected database and S-Matrix. If length-1 in projected database fulfills minimum support then S-Matrix for length-2 will be build. If S-Matrix fulfills minimum support, S-matrix will be stored and re-process. The process will be stop if the number of projected database less then minimum support. S-Matrix which fulfills minimum support will be place in the end of length-x and the result of length-1 from projected database will be place in the end of length-x. If all the recurrent process is finished, length-x Sequential Pattern is created.

IV. Discovering Sequential Disease Patterns Testing

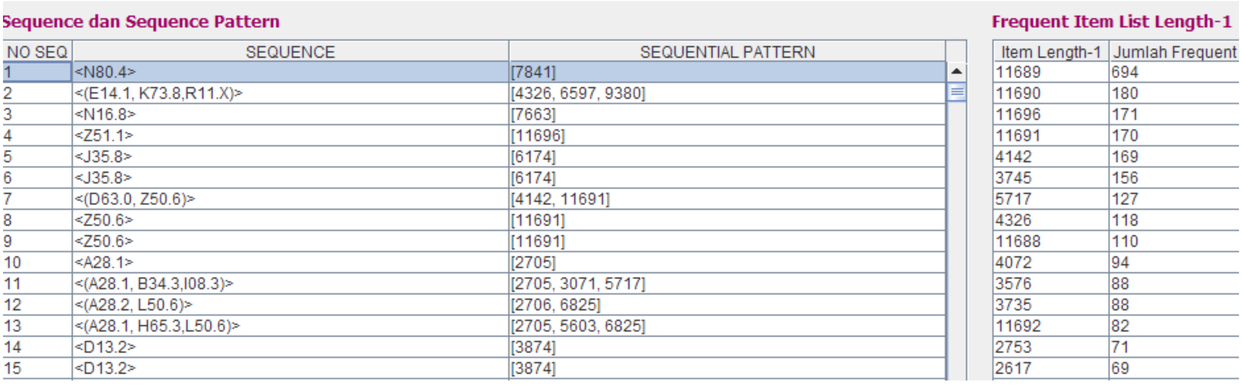
Testing is using data from January 1, 2003 until December 31, 2003, with minimum support = 2, time period = 6 days, with all province (Fig. 2).

Sequence and sequence pattern is built from data in Fig. 1 based on sequence number. Fig. 3 is shown frequent item list length-1 which fulfills minimum support and Fig. 4 is shown the triangular matrix. Triangular matrix which fulfills minimum support becomes length-2 sequential pattern.



Keterangan	WKTMSK	WKTKLR	NO_SEQ
N81.3	2003-01-05	2003-01-14	1
E14.9	2003-01-07	2003-01-15	2
K74.6	2003-01-07	2003-01-15	2
R18.X	2003-01-07	2003-01-15	2
N18.9	2003-05-19	2003-05-22	3
Z51.9	2003-07-02	2003-07-11	4
J38.3	2003-05-04	2003-05-06	5
J38.3	2003-11-02	2003-11-04	6
D64.9	2003-06-02	2003-06-06	7
Z51.3	2003-06-02	2003-06-06	7
Z51.3	2003-09-19	2003-09-23	8
Z51.3	2003-11-10	2003-11-17	9
A30.4	2003-01-28	2003-01-31	10
A30.4	2003-06-09	2003-06-19	11
B35.4	2003-06-09	2003-06-19	11
I10.X	2003-06-09	2003-06-19	11
A30.5	2003-09-11	2003-09-17	12
L52.X	2003-09-11	2003-09-17	12

Fig. 2. Data January 1, 2003 until December 31, 2003



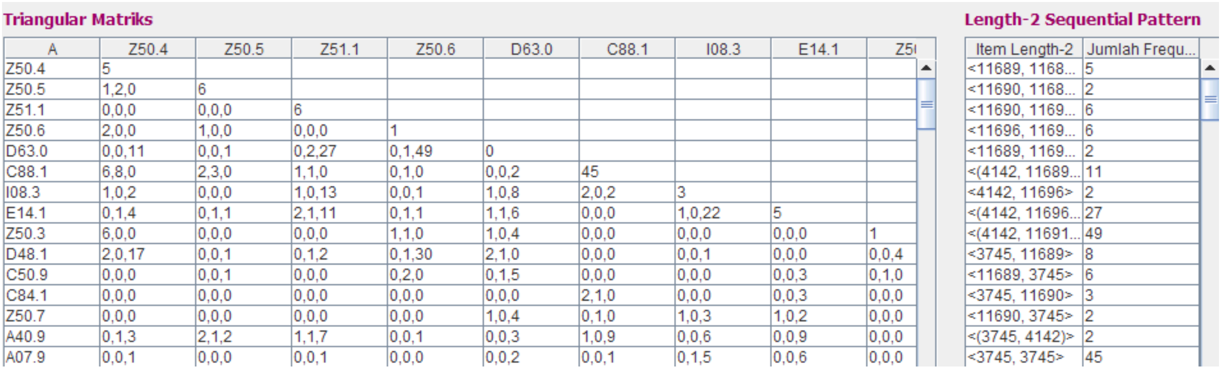
Sequence dan Sequence Pattern

NO SEQ	SEQUENCE	SEQUENTIAL PATTERN
1	<N80.4>	[7841]
2	<(E14.1, K73.8,R11.X)>	[4326, 6597, 9380]
3	<N16.8>	[7663]
4	<Z51.1>	[11696]
5	<J35.8>	[6174]
6	<J35.8>	[6174]
7	<(D63.0, Z50.6)>	[4142, 11691]
8	<Z50.6>	[11691]
9	<Z50.6>	[11691]
10	<A28.1>	[2705]
11	<(A28.1, B34.3,I08.3)>	[2705, 3071, 5717]
12	<(A28.2, L50.6)>	[2706, 6825]
13	<(A28.1, H65.3,L50.6)>	[2705, 5603, 6825]
14	<D13.2>	[3874]
15	<D13.2>	[3874]

Frequent Item List Length-1

Item Length-1	Jumlah Frequent
11689	694
11690	180
11696	171
11691	170
4142	169
3745	156
5717	127
4326	118
11688	110
4072	94
3576	88
3735	88
11692	82
2753	71
2617	69

Fig. 3. Sequence and sequence pattern (left), frequent item list length-1 (right)



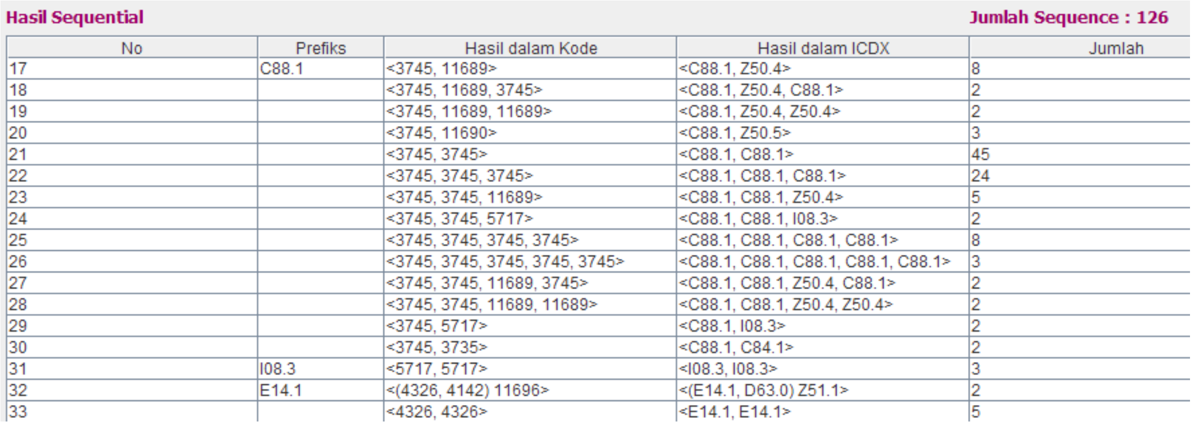
Triangular Matriks

A	Z50.4	Z50.5	Z51.1	Z50.6	D63.0	C88.1	I08.3	E14.1	Z5
Z50.4	5								
Z50.5	1,2,0	6							
Z51.1	0,0,0	0,0,0	6						
Z50.6	2,0,0	1,0,0	0,0,0	1					
D63.0	0,0,11	0,0,1	0,2,27	0,1,49	0				
C88.1	6,8,0	2,3,0	1,1,0	0,1,0	0,0,2	45			
I08.3	1,0,2	0,0,0	1,0,13	0,0,1	1,0,8	2,0,2	3		
E14.1	0,1,4	0,1,1	2,1,11	0,1,1	1,1,6	0,0,0	1,0,22	5	
Z50.3	6,0,0	0,0,0	0,0,0	1,1,0	1,0,4	0,0,0	0,0,0	0,0,0	1
D48.1	2,0,17	0,0,1	0,1,2	0,1,30	2,1,0	0,0,0	0,0,1	0,0,0	0,0,4
C50.9	0,0,0	0,0,1	0,0,0	0,2,0	0,1,5	0,0,0	0,0,0	0,0,3	0,1,0
C84.1	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	2,1,0	0,0,0	0,0,3	0,0,0
Z50.7	0,0,0	0,0,0	0,0,0	0,0,0	1,0,4	0,1,0	1,0,3	1,0,2	0,0,0
A40.9	0,1,3	2,1,2	1,1,7	0,0,1	0,0,3	1,0,9	0,0,6	0,0,9	0,0,0
A07.9	0,0,1	0,0,0	0,0,1	0,0,0	0,0,2	0,0,1	0,1,5	0,0,6	0,0,0

Length-2 Sequential Pattern

Item Length-2	Jumlah Frequ...
<11689, 1168...	5
<11690, 1168...	2
<11690, 1169...	6
<11696, 1169...	6
<11689, 1169...	2
<(4142, 11689...	11
<4142, 11696>	2
<(4142, 11696...	27
<(4142, 11691...	49
<3745, 11689>	8
<11689, 3745>	6
<3745, 11690>	3
<11690, 3745>	2
<(3745, 4142)>	2
<3745, 3745>	45

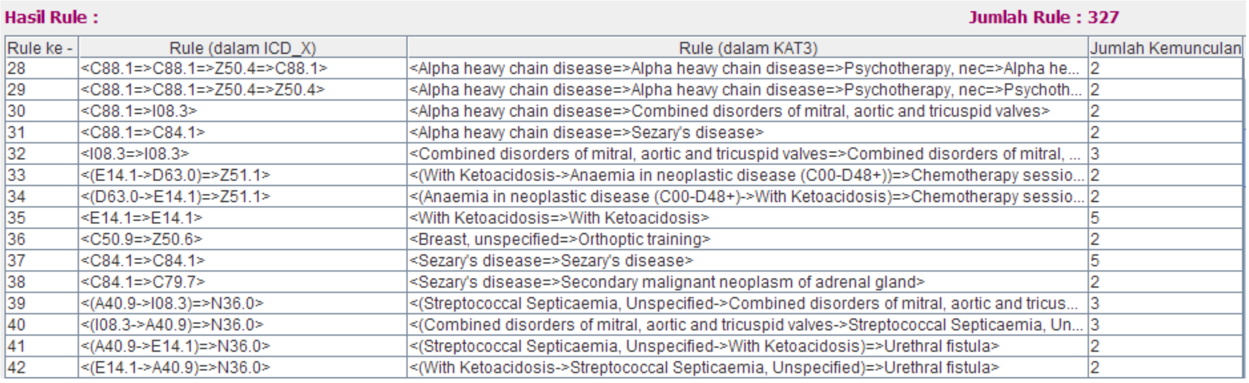
Fig. 4. Triangular matrix (left), length-2 sequential pattern (right)



Hasil Sequential **Jumlah Sequence : 126**

No	Prefiks	Hasil dalam Kode	Hasil dalam ICDX	Jumlah
17	C88.1	<3745, 11689>	<C88.1, Z50.4>	8
18		<3745, 11689, 3745>	<C88.1, Z50.4, C88.1>	2
19		<3745, 11689, 11689>	<C88.1, Z50.4, Z50.4>	2
20		<3745, 11690>	<C88.1, Z50.5>	3
21		<3745, 3745>	<C88.1, C88.1>	45
22		<3745, 3745, 3745>	<C88.1, C88.1, C88.1>	24
23		<3745, 3745, 11689>	<C88.1, C88.1, Z50.4>	5
24		<3745, 3745, 5717>	<C88.1, C88.1, I08.3>	2
25		<3745, 3745, 3745, 3745>	<C88.1, C88.1, C88.1, C88.1>	8
26		<3745, 3745, 3745, 3745, 3745>	<C88.1, C88.1, C88.1, C88.1, C88.1>	3
27		<3745, 3745, 11689, 3745>	<C88.1, C88.1, Z50.4, C88.1>	2
28		<3745, 3745, 11689, 11689>	<C88.1, C88.1, Z50.4, Z50.4>	2
29		<3745, 5717>	<C88.1, I08.3>	2
30		<3745, 3735>	<C88.1, C84.1>	2
31	I08.3	<5717, 5717>	<I08.3, I08.3>	3
32	E14.1	<(4326, 4142) 11696>	<(E14.1, D63.0) Z51.1>	2
33		<4326, 4326>	<E14.1, E14.1>	5

Fig. 5. Sequential pattern



Hasil Rule : **Jumlah Rule : 327**

Rule ke -	Rule (dalam ICD_X)	Rule (dalam KAT3)	Jumlah Kemunculan
28	<C88.1=>C88.1=>Z50.4=>C88.1>	<Alpha heavy chain disease=>Alpha heavy chain disease=>Psychotherapy, nec=>Alpha he...	2
29	<C88.1=>C88.1=>Z50.4=>Z50.4>	<Alpha heavy chain disease=>Alpha heavy chain disease=>Psychotherapy, nec=>Psychoth...	2
30	<C88.1=>I08.3>	<Alpha heavy chain disease=>Combined disorders of mitral, aortic and tricuspid valves>	2
31	<C88.1=>C84.1>	<Alpha heavy chain disease=>Sezary's disease>	2
32	<I08.3=>I08.3>	<Combined disorders of mitral, aortic and tricuspid valves=>Combined disorders of mitral, ...	3
33	<(E14.1->D63.0)=>Z51.1>	<(With Ketoacidosis->Anaemia in neoplastic disease (C00-D48+))=>Chemotherapy sessio...	2
34	<(D63.0->E14.1)=>Z51.1>	<(Anaemia in neoplastic disease (C00-D48+)->With Ketoacidosis)=>Chemotherapy sessio...	2
35	<E14.1=>E14.1>	<With Ketoacidosis=>With Ketoacidosis>	5
36	<C50.9=>Z50.6>	<Breast, unspecified=>Orthoptic training>	2
37	<C84.1=>C84.1>	<Sezary's disease=>Sezary's disease>	5
38	<C84.1=>C79.7>	<Sezary's disease=>Secondary malignant neoplasm of adrenal gland>	2
39	<(A40.9->I08.3)=>N36.0>	<(Streptococcal Septicaemia, Unspecified->Combined disorders of mitral, aortic and tricus...	3
40	<(I08.3->A40.9)=>N36.0>	<(Combined disorders of mitral, aortic and tricuspid valves->Streptococcal Septicaemia, Un...	3
41	<(A40.9->E14.1)=>N36.0>	<(Streptococcal Septicaemia, Unspecified->With Ketoacidosis)=>Urethral fistula>	2
42	<(E14.1->A40.9)=>N36.0>	<(With Ketoacidosis->Streptococcal Septicaemia, Unspecified)=>Urethral fistula>	2

Fig. 6. Rules of the sequential

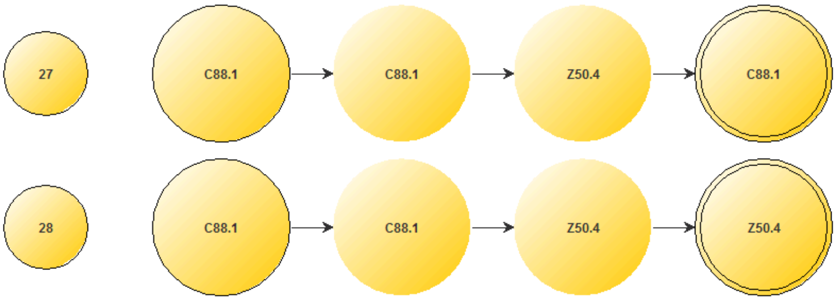


Fig. 7. Graph for rule 27th and 28th

Fig. 5 is the result of *length-x* process in code form and Fig. 6 is the result with the diagnosis's name. For example, alpha heavy chain disease is followed by psychotherapy nec. Sequential pattern can be figured in the form of automata graph. Fig. 7 is representing sequence number 27 and 28 of Fig. 6. Single solid circle represents start state and double and double solid circle represents end state. Circle with no border represents transition state.

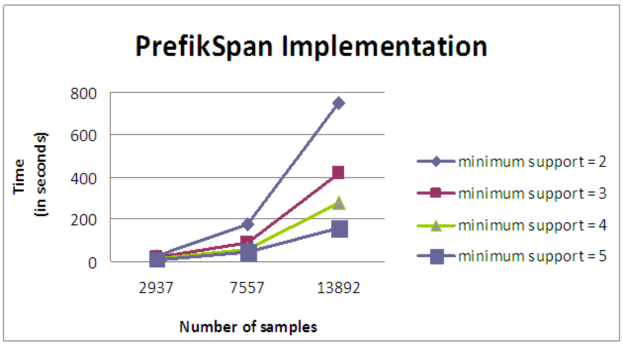


Fig. 8. Graphic of PrefixSpan implementation

This research also compares processing speed between PrefixSpan, FreeSpan and Spade method. The comparison with different minimum support shows the same result (Fig. 9 amd Fig. 10). Different amount of data shows processing data with FreeSpan method is the fastest. While Spade method shows the speeding process with larger data is not efficient.

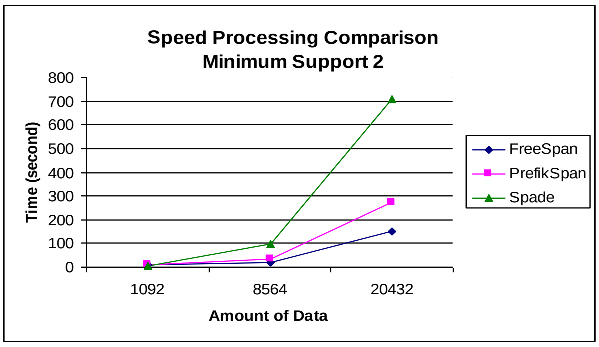


Fig. 9. Comparison between PrefikSpan method, FreeSpan method with minimum support 2

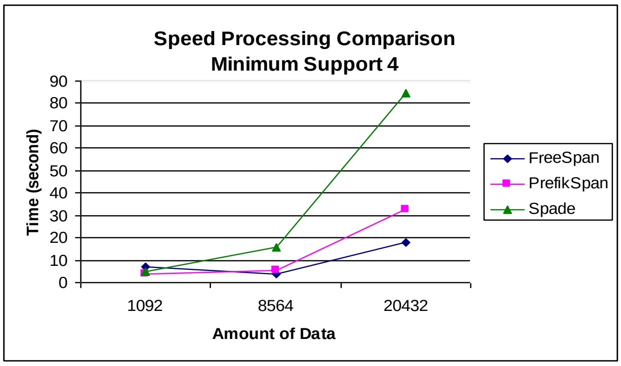


Fig. 10. Comparison between PrefikSpan method, FreeSpan method with minimum support 4

V. Conclusion

PrefixSpan method can be used for mining sequential diseases pattern from database sequential. The generated patterns can be used a knowledge to predict sequential diseases. As a result the medical representative can take preventive and curative action more precisely. The rule will increase as the smaller *minimum support*, but it will cost time processing.

Acknowledgment

We would like to thank Directorate General of Indonesian Higher Education as sponsor and financial support to implement this work.

References

[1] Indonesia, Departemen Kesehatan, *Sistem informasi rumah sakit di Indonesia, Jakarta*, 2000.

[2] J. Han and M. Kamber, "Data mining: Concept and techniques," Canada: Simon Fraser University, 2000.

[3] J. Han *et al*, "Mining sequential patterns by pattern-growth: The PrefixSpan approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, pp. 1424-1140, November 2004.

[4] J. Han *et al*, "PefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth," in *Proc. 17th International Conference on Data Engineering, German*, pp. 215-224, 2001.

[5] Kusuma, R. Yinsi, and R. Yapola, "Implementation of star schema and data warehouse at RSU Dr. Soetomo," Petra Christian University, Surabaya, 2009.

Silvia Rostianingsih was born in Indonesia. She earned her Master and Bachelor's degree from Institute Technology of Sepuluh Nopember Surabaya. She has taught at Petra Christian University since 2001. Her researches focused on information system and database. She also becomes a member of IACSIT.

Gregorius Satia Budhi was born in Indonesia. He earned his Master's degree from Institute Technology of Sepuluh Nopember Surabaya. He earned his Bachelor's degree from Institute Technology of Adhi Tama Surabaya. He has taught at Petra Christian University since 2002. His researches focused on data mining and artificial intelligent.

Leonita Kumalasari Theresia was born in Indonesia. She earned her Bachelor's degree from Petra Christian University. Now she is taking her Master's degree at University of Surabaya. She also works as a IT consultant.