ISSN: 2089-1121

WEB USAGE MINING BERBASIS WAP-TREE

Lukman Zaman*), Joan Santoso), Rudy Wijaya**),**
dan Jessica Felani Wijoyo)**
*) Program Pascasarjana Teknologi Informasi
Sekolah Tinggi Teknik Surabaya
**) Jurusan Teknik Informatika
Sekolah Tinggi Teknik Surabaya
luqmanz@gmail.com , joan@stts.edu , - , jessicafelaniwijoyo@gmail.com

ABSTRAK

Pada saat ini, pertumbuhan data pada *World Wide Web* semakin besar, sehingga analisa dan penemuan informasi yang berguna dari World Wide Web menjadi suatu kebutuhan yang penting. *Web access pattern*, yang merupakan suatu urutan akses yang sering dilakukan oleh user dalam suatu web site, adalah suatu knowledge (pengetahuan) yang menarik dan berguna.Web access pattern juga dikenal sebagai sequential pattern mining dalam suatu file Web log (*Web log mining*).

Sequential pattern mining, yang merupakan proses penemuan pola yang sering muncul dalam suatu database sequence, pertama kali diperkenalkan oleh Agrawal and Srikant. Konsep dari sequential pattern mining adalah sebagai berikut: diberikan suatu database sequence dimana tiap sequence adalah suatu daftar akses user yang diurutkan berdasarkan waktu akses dan tiap akses terdiri dari kumpulan informasi yang berguna, kemudian dicari semua pola akses dengan minimum support yang didefinisikan oleh user, dimana support merupakan jumlah dari database sequence yang mengandung pola tersebut.

Pada penelitian ini dibahas mengenai suatu cara mining web access pattern dari web log secara efisien. Suatu struktur data baru, yang disebut sebagai *Web Access Pattern tree*, atau disingkat WAP-tree, digunakan untuk mining web access pattern dari web log secara efisien. Untuk menunjukkan kelebihan dari algoritma WAP-tree, digunakan algoritma GSP sebagai pembanding, dimana algoritma GSP merupakan perkembangan dari Apriori All..

Kata kunci :Web Mining, WAP-tree, GSP (Generalized Sequential Patterns)

ABSTRACT

With the explosive growth of data available on the World Wide Web, discovery and analysis of useful information from the World Wide Web becomes a practical necessity. Web access pattern, which is the sequence of accesses pursued by users frequently, is a kind of interesting and useful knowledge in practice. Web access pattern is also known as a sequential pattern mining in a large set of pieces of Web logs (Web log mining).

Sequential pattern mining, which discovers frequent patterns in a sequence database, was first introduced by Rakesh Agrawal and Ramakrishnan Srikant as follows: a sequence database is given where each sequence is a list of users accesses ordered by access time and each access consists of a set of information, all sequential

ISSN: 2089-1121

patterns are found with a user-specified minimum support, where the support is the number of data sequences that contain the pattern.

In this research, the problem of mining access pattern from Web logs efficiently is studied. A novel data structure, called Web access pattern tree, or WAP-tree in short, is developed for efficient mining of access patterns from pieces of logs. The Web access pattern tree stores highly compressed, critical information for access pattern mining and facilitates the development of novel algorithms for mining access patterns in large set of log pieces. To show the performance of WAP-tree algorithm, GSP algorithm, which is an Apriori All-based algorithm, is used as a comparison.

Keywords :Web Mining, WAP-tree, GSP (Generalized Sequential Patterns)

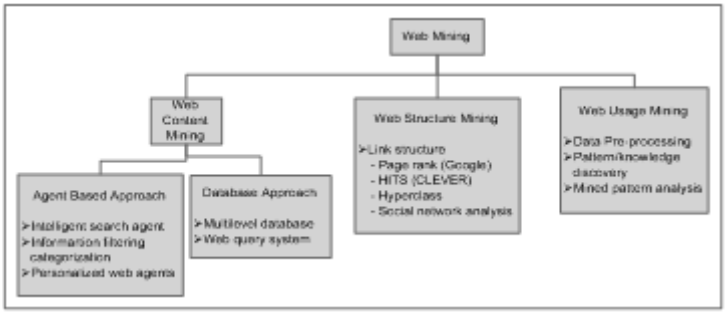
1. PENDAHULUAN

Pada saat ini, pertumbuhan data pada World Wide Web semakin besar, sehingga analisa dan penemuan informasi yang berguna dari World Wide Web menjadi suatu kebutuhan yang penting. Web access pattern, yang merupakan suatu urutan akses yang sering dilakukan oleh user dalam suatu web site, adalah suatu knowledge (pengetahuan) yang menarik dan berguna.Web access pattern juga dikenal sebagai sequential pattern mining dalam suatu file Web log (Web log mining).

Pada penelitian ini dibahas mengenai suatu cara mining web access pattern dari web log secara efisien. Suatu struktur data baru, yang disebut sebagai Web Access Pattern tree, atau disingkat WAP-tree, digunakan untuk mining web access pattern dari web log secara efisien.Untuk menunjukkan kelebihan dari algoritma WAP-tree, digunakan algoritma GSP sebagai pembanding, dimana algoritma GSP merupakan perkembangan dari Apriori All..

2. TINJAUAN PUSTAKA

Web mining telah didefinisikan sebagai sebuah aplikasi atau bagian dari teknologi data mining yang dikhususkan untuk menanggulangi data WWW (World Wide Web) yang sangat besar.Berdasarkan prosesnya, web mining dapat dibagi menjadi tiga kategori umum (dapat dilihat pada Gambar 1) antara lain *Web contents mining*, *Web structure mining*, dan *Web usage mining.*

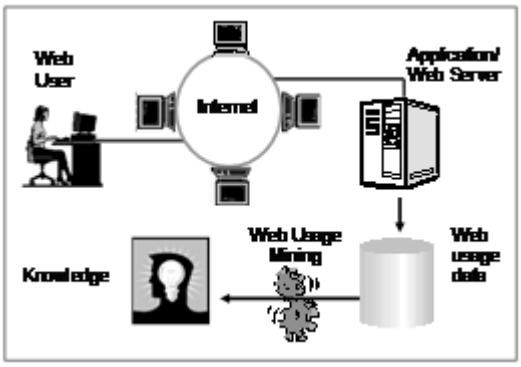


Gambar 1. Taxonomy Web Mining

Web usage mining merupakan sebuah bidang penelitian yang berfokus pada perkembangan teknik dan tool untuk mempelajari kebiasaan user dalam web. Web usage mining juga dapat dikatakan sebagai web log mining (suatu proses untuk menemukan pola-pola yang berguna dalam web access log). Organisasi – organisasi biasanya mengumpulkan data operasi sehari-harinya dalam jumlah besar yang secara

ISSN: 2089-1121

otomatis di-generate oleh web server. Data tersebut disimpan dalam web server access log.Secara umum proses dari web usage mining ini dapat dilihat pada gambar 2.



Gambar 2. Proses Umum Web Usage Mining

3. METODE PENELITIAN

Dalam pelaksanaan penelitian ini dibutuhkan beberapa tahap proses yang harus dilakukan. Tahapan-tahapan yang dilakukan adalah sebagai berikut:

- Melakukan preprocessing data agar data yang ada dapat secara mudah untuk dilakukan pencarian/ekstraksi pengetahuan.
- Melakukan mining pada data dengan menggunakan algoritma WAP-Tree.
- Melakukan mining pada data dengan menggunakan algoritma GSP
- Evaluasi hasil dari algoritma WAP-Tree dan GSP.
- Membandingkan efisiensi algoritma WAP-Tree dan GSP.

4. ALGORITMA WAP-TREE

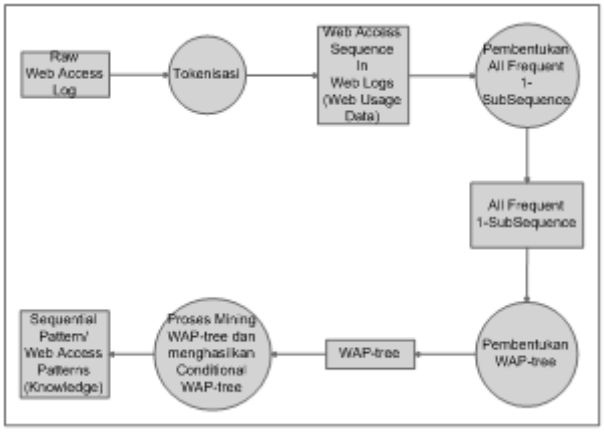
Pada dasarnya web access pattern merupakan suatu pola akses pengunjung web yang terkandung di dalam suatu web access log, dimana pola akses tersebut telah memenuhi minimum support yang didefinisikan. Frequent event adalah suatu event yang telah memenuhi minimum support, sedangkan support untuk suatu event adalah jumlah *user/client* dalam web access sequence database yang mengandung *event* tersebut. Saat ini terdapat banyak aplikasi yang mengarah ke bidang ekstraksi suatu pola/pattern dari web log.

Dalam kenyataannya, proses ekstraksi web access pattern dari web access log memperoleh banyak pengetahuan yang berguna dan menarik, pengetahuan tersebut dapat berupa sequential pattern yang kebanyakan merupakan perkembangan dari teknik association rule mining. Manfaat dari dikembangkannya aplikasi web log mining/web usage mining ini adalah untuk meningkatkan design sebuah web site, menganalisa performance suatu system, mengetahui tingkah-laku user/client beserta tujuannya (pola akses pengunjung web/web access pattern), dan membantu dalam penyusunan sebuah adaptive web site.

Secara umum proses yang dilakukan dalam algoritma WAP-tree ini yaitu melakukan pembacaan web access sequence database sebanyak dua kali kemudian secara rekursif melakukan proses mining sebuah tree yang telah dibentuk pada saat pembacaan kedua dengan menggunakan *conditional search*. Pada saat pembacaan web access sequence database yang pertama bertujuan untuk membentuk kumpulan event yang memenuhi support atau dengan kata lain pembentukan frequent

ISSN: 2089-1121

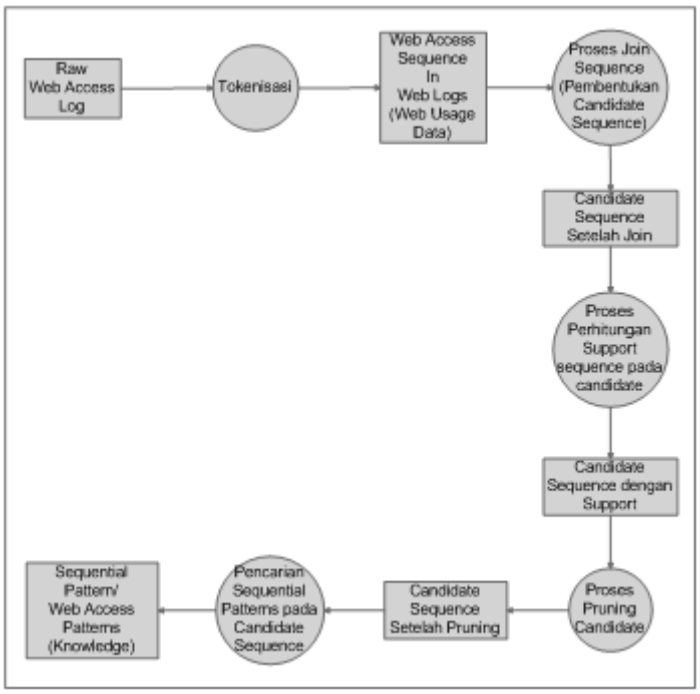
subsequence.Sedangkan pada saat pembacaan web access sequence database yang kedua kalinya, digunakan untuk membentuk struktur data berupa sebuah tree atau yang biasa dikenal sebagai struktur data WAP-tree. Alur proses dari algoritma WAP-tree dapat dilihat dalam gambar 3.



Gambar 3. Alur Proses AlgoritmaWAP-tree

5. ALGORITMA GSP

Beberapa algoritma telah banyak dikemukakan untuk menyelesaikan masalah sequential pattern mining. Algoritma GSP merupakan salah satu algoritma konvensional yang telah ada. Algoritma GSP diperkenalkan oleh Ramakrishnan Srikant dan Rakesh Agrawal. Algoritma GSP merupakan perkembangan dari algoritma Apriori All. Secara garis besar algoritma GSP dapat dibagi menjadi beberapa tahap, yaitutahap pembentukan candidate sequence, tahap perhitungan support dan proses pruning, dan tahap pencarian seluruh sequential pattern yang ada.Alur proses pada algoritma GSP dapat dilihat pada gambar 4.



Gambar 4. Alur Proses AlgoritmaGSP

ISSN: 2089-1121

Berdasarkan alur proses dalam gambar 4, tahap pembentukan candidate sequence ini dapat dibagi menjadi dua bagian utama. Kedua bagian tersebut adalah sebagai berikut :

a. Proses Join.
   Proses join yang dilakukan dalam algoritma GSP ini bertujuan untuk membentuk candidate *k*-sequence dengan cara melakukan penggabungan atau join antara(*k*-1)-sequence dengan (*k*-1)-sequence. Suatu sequence S_1 ((*k*-1)-sequence) dapat dijoin dengan sequence S_2 ((*k*-1)-sequence), jika subsequence dari S_1 yang diperoleh dengan menghilangkan event pertama dari sequence S_1 adalah sama dengan subsequence yang diperoleh dengan menghilangkan event terakhir dari sequence S_2.
b. Proses Prune.
   Pada proses prune dalam algoritma GSP berguna untuk mengurangi jumlah sequence dalam sebuah candidate *k*-sequence yang biasanya memiliki ukuran relatif besar. Sequence yang biasanya dihilangkan atau dipruning adalah sequence yang tidak memenuhi minimum support.

6. EVALUASI

Untuk keperluan eksperimen, maka akan digunakan tiga buah file database sequence berupa file web log. File database atau file web log tersebut disediakan oleh Mike Perkowitz dan Oren Etzioni berikut dengan keterangan untuk format dari masing-masing file Web Log. File database sequence atau file web log yang akan dipakai untuk percobaan adalah sebagai berikut :

a. File Web Log m.981123.
   File web log m.981123 memiliki format yang standart. File web log ini mencatat pengaksesan userpada website "music machines" (http://machines.hyperreal.org). Contoh segmen dari file web log ini adalah sebagai berikut :

   ```
   www.hyperreal.org|anon00000000000000101487|
    GET/music/machines/
   manufacturers/ HTTP/1.1|text/html|200|1998/11/22-23:59:56|-|6316|-|-|
    http://www.hyperreal.org/
   music/machines/ manufacturers/Roland/MC-303/|Mozilla/4.0 (compatible; MSIE 4.0; Windows 95)
   ```

b. File Web Log m.970418.
   File web log m.970418 ini sama dengan file web log m.981123. Akan tetapi file web log ini memiliki format yang kurang standart. Contoh segmen dari file web log ini adalah sebagai berikut :

   ```
   O:00000000000000057273 || T:1997/04/18-13:40:19 || U:/map.html ||
     R:http://www.hyperreal.com/machines/categories/       images/bullet.gif?MMAgent
   ```

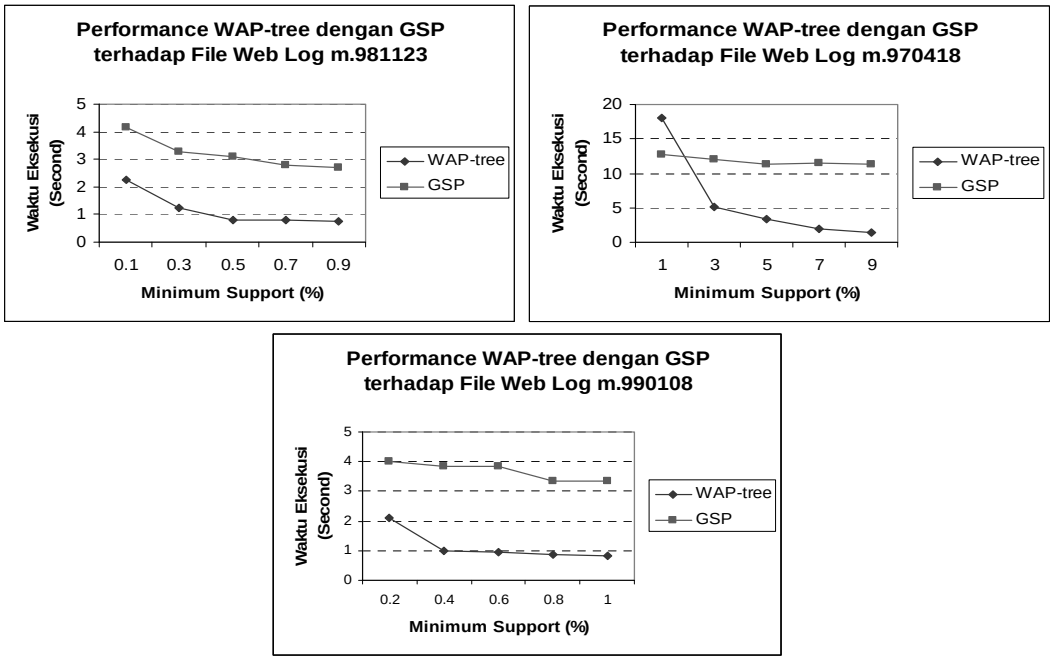
c. File Web Log m.990108.
   File web log ini memiliki format yang sama dengan file web log m.981123. File web log ini juga mencatat pengaksesan user pada website "music machines". File web log ini terdapat 16374 user. Contoh segmen dari file web log ini adalah sebagai berikut :

   ```
   www.hyperreal.org|anon00000000000000157746|GET /music/machines/categories/drum-machines/samples/
     HTTP/1.1|text/html|200|1999/01/07-23:59:17|-|-|-|-
     |http://www.hyperreal.org/music/machines/samples.html?MMAgent|Mozilla/4.0 (compatible; MSIE 4.01; Windows
     95)
   ```

Pada beberapa percobaan yang dilakukan, diberikan beberapa nilai minimum support yang bervariasi. Berikut ini adalah hasil evaluasi dari Algoritma WAP-Tree dan GSP yang berada pada gambar 5. Percobaan ini bertujuan untuk melihat seberapa efisien

ISSN: 2089-1121

algoritma WAP-tree ini bila dibandingkan dengan algoritma konvensional yaitu algoritma GSP.



Gambar 5. Perbandingan WAP-tree Dengan GSP

7. PENUTUP

Dari hasil penelitian yang dilakukan, maka dapat diambil beberapa kesimpulan sebagai berikut.Dalam menyelesaikan masalah web log mining (sequential pattern mining), dapat dilihat bahwa algoritma WAP-tree lebih efisien bila dibandingkan dengan algoritma GSP (Generalized Sequential Patterns). Hal ini disebabkan karena untuk mencari seluruh web access pattern yang ada dalam suatu web log yang menggunakan algoritma WAP-tree ini tidak memerlukan pembentukan candidate sequence, dimana proses pembentukan candidate sequence merupakan salah satu inti dari algoritma GSP. Waktu eksekusi yang dibutuhkan oleh algoritma WAP-tree lebih sedikit bila dibandingkan dengan algoritma GSP.Ditinjau dari keseluruhan proses kedua algoritma tersebut dapat dikatakan bahwa algoritma WAP-tree hanya membutuhkan pembacaan atau scan web access sequence database sebanyak dua kali. Sedangkan pembacaan atau scan WAS database dalam algoritma GSP adalah sejumlah maximal dari panjang frequent sequence yang ada (untuk keseluruhan proses pembacaan akan dilakukan berkali-kali). Tree yang dihasilkan oleh algoritma WAP-tree sangat efisien karena tree tersebut memiliki sifat yaitu : height dari tree adalah satu ditambah dengan panjang maximal sequence dalam file web log dan ukuran yang dihasilkan juga jauh lebih kecil dari ukuran file web log. Hal ini disebabkan karena adanya teknik *prefix sharing* (sequence yang memiliki prefix sequence yang sama akan ditampung ke dalam satu cabang dalam tree).

Dengan adanya perkembangan dibidang sequential pattern mining khususnya web log mining menyebabkan banyak bermunculan algoritma baru yang menyatakan memiliki performance yang lebih baik jika dibandingkan dengan algoritma yang telah ada (GSP). Ada beberapa saran yang dapat diberikan untuk perkembangan pembahasan algoritma sequential pattern mining yang lain adalah sebagai berikut. Algoritma CS-

ISSN: 2089-1121

Mine, Dasar dari algoritma ini adalah sama pembentukan sebuah tree, dimana tree tersebut biasanya disebut dengan WAP-tree. Dapat dikatakan bahwa algoritma CS-Mine ini merupakan perkembangan yang lebih efisien dari algoritma WAP-tree.Algoritma PrefixSpan, algoritma ini merupakan perkembangan dari algoritma GSP. Dalam menyelesaikan masalah sequential pattern mining, algoritma ini menggunakan metode pattern-growth. Inti dari algoritma ini juga hampir sama dengan algoritma GSP yaitu terletak pada pembentukan candidate sequence.

8. DAFTAR PUSTAKA

Agrawal, R., Srikant,R.. *Fast Algorithms for Mining Association Rules*. In Proceedings of International Conference on Very Large Data Bases. 1994.

Agrawal, R., Srikant,R..*Mining sequential patterns*.In Proc. 1995 Int. Conf. Data Engineering, Taipei, Taiwan. 1995.

Agrawal, R., Srikant,R..*Mining Sequential Patterns: Generalizations and Performance Improvements*. IBM Research Report RJ9994, IBM Almaden Research Center.1995.

Antunes. C,Oliveira,L..*A. Generalization of Pattern-growth Methods for Sequential Pattern Mining with Gap Constraints*.

Mobasher,B., Jain,N., Han,E.H., Srivastava,J..*Web mining : Pattern Discovery from World Wide Web Transactions*. Technical Report TR96-050, Department of Computer Science, University of Minnesota.1996.

Mobasher,B., Cooley,R., Srivastava J.. *Web mining: Information and pattern discovery on the world wide web.* In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Paris, France.1997.

Degan., Q, Zhongtao., Z, Paterno.,M.C. S. *Web Usage Mining*.2001.

Galeas., P. *Web Mining*. http://www.galeas.de/webmining.html

Pei,J., Han,J., Mortazavi-Asl,B., Zhu,H..*Mining Access Patterns Efficiently from Web Logs*.In Proc. 2000 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'00), Kyoto, Japan.2000.

Borges,Jose Luis Cabral de Moura. *A Data Mining Model Capture User Web Navigation Pattern*.2000.

Kumar., V. *Discovery of Indirect Association from Web Usage Data*. 2002.

Page., L, Brin., S, Motwani., R, Winograd., T.*The PageRank Citation Ranking: Bringing Order to the Web*. 1998.

Perkowitz, M., Etzioni,O..*Adaptive Sites : Automatically Learning from User Access patterns*. In Proc. 6th Int'l World Wide Web Conf, Santa Clara, California.1995

W3C. *Extended Log File Format*.http://www.w3.org/protocols

Wei, G. *An Overview on Web Mining*.2000.

Yue-Shi., L.*Advanced Topics in Data Mining : Web Mining*.

Zaiane.Osmar R, Xin,M., Han,J. *Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs*.In Proc. Advances inDigital Libraries Conf. (ADL'98), Melbourne, Australia..1998.

Zaiane, Osmar R..*Resource and Knowledge Discovery from The Internet and Multimedia Repositories*. PhD thesis, Computing Science, Simon Fraser University.1999.