Pencarian Asosiasi Topik Dalam Ayat Al Qur'an
Dengan Menerapkan Algoritma *Multipass Direct Hashing and Prunning* (M-DHP)

***Alfian Ardhi*[1)], *Lailil Muflikhah*[2)], *Marji*[3)]**
Program Studi Ilmu Komputer, Program Teknologi Informasi dan Ilmu Komputer
Universitas Brawijaya, Jl. Veteran Malang, 65145, Indonesia
E-mail : [1] alfianardhimy@gmail.com, [2]laililmf@gmail.com, [3]marji@ub.ac.id

ABSTRAK

Seluruh umat islam di seluruh dunia dan tidak terkecuali di Indonesia pasti memiliki keinginan untuk lebih memperdalam pengetahuan yang terkandung di dalam Al Quran, dimana ketika seseorang memperdalam satu topik tertentu maka besar kemungkinan topik tersebut memiliki keterkaitan dengan topik-topik lain yang terdapat pada Al Qur'an. Oleh karena itu, pada penelitian ini dimaksudkan untuk membuat sistem pencarian asosiasi atau keterkaitan topik pada Al Qur'an. Untuk metode yang digunakan pada penelitian Association Rule Mining ini adalah Multipass Direct Hashing and Prunning (M-DHP), dimana metode ini untuk mengatasi karakteristik dari text database yang memerlukan ruang memory besar ketika menghitung frequent itemset. Proses pembangkitan rule pada M-DHP ini diperoleh melalui pengolahan data transaki yang terbentuk dan kemudian menemukan frequent 1 itemset. Dari frequent 1 itemset ini dilakukan proses partisi yang digunakan untuk menemukan frequent k itemset. Selama proses pembentukan frequent k itemset, secara bersamaan juga melakukan reduksi database transaksi. Setelah seluruh frequent itemset terbentuk, maka tahapan selanjutnya merupakan pembangkitan rule yang digunakan untuk menemukan pola asosiasi topik pada al qur'an. Pengujian rule pada penelitian ini menggunakan rule-rule dengan tingkat akurasi yang kuat berdasarkan pada conviction dan hiper lift ratio, untuk partisi dipilih pada partisi 4. Pemilihan k-partisi yang tidak telalu besar memiliki keunggulan yaitu semakin banyak kombinasi rule yang dihasilkan. Hasil pengujian terbaik ketika minimum support 8% dan minimum confidence 90% dengan hasil 81,8%.

***Kata Kunci** : Association Rule Mining, M-DHP, Al-Qur'an, frequent itemset*

1. PENDAHULUAN

Al Qur'an merupakan kitab suci agama Islam sehingga dapat dipastikan bahwa seluruh umat atau pemeluk agama Islam banyak yang memperdalam untuk mempelajari berbagai pengetahuan yang terkandung pada ayat-ayat Al Qur'an. Saat ini untuk lebih memperdalam Al Qur'an masyarakat dapat mempelajari dengan memanfaatkan Al Qur'an terjemahan baik dalam bentuk fisik maupun Al Qur'an digital. Pada Al Qur'an terjemahan baik yang dalam bentuk fisik ataupun digital telah dikelompokkan berdasarkan topik dari ayat tersebut, namun belum ada yang menampilkan suatu bentuk keterkaitan atau asosiasi antar topik, padahal berbagai topik yang terkandung di dalam Al Qur'an akan saling terkait satu sama lain.

Apriori adalah algoritma pertama untuk aturan asosiasi yang diusulkan oleh Agrawal dan Srikant, dimana digunakan untuk menemukan kombinasi item dengan berdasar pada aturan tertentu kemudian diuji apakah kombinasi item tersebut memenuhi syarat *Support* Minimum *(Minimum Support)*. Kombinasi item yang memenuhi syarat *Support* minimum tersebut disebut *frequent itemeset* dimana nanti dipergunakan untuk membentuk aturan-aturan yang diharapkan memenuhi syarat minimum *Confident* [1].

Algoritma Apriori memiliki kelemahan yaitu harus membangkitkan kandidat

Itemset dalam jumlah besar, oleh karena itu algoritma Apriori dikembangkan agar lebih efisien dan salah satunya adalah algoritma DHP(*Direct Hashing and Prunning*). Algoritma DHP ini melakukan dua proses untuk menyelesaikan permasalahn tersebut. Pertama menggunakan *Hash Table* untuk mengurangi kandidat *itemset*, sedangkan proses kedua dengan melakukan proses reduksi pada database transaksi [6].

Menurut John dkk [4], kedua algoritma tersebut yaitu Apriori dan *Direct Hashing and Prunning* (DHP) tidak cocok untuk menemukan *frequent itemset* pada data transaksi yang terbentuk dari *text database*. Hal ini disebabkan untuk menghitung seluruh kejadian sehingga diperoleh jumlah *large* setiap *frequent itemset* dari kombinasi item (*itemset*) yang terbentuk memerlukan *space memory* yang cukup besar, oleh karena itu algoritma *Direct Hashing and Prunning* (DHP) dikembangkan agar lebih efisien menjadi *Multipass Direct Hashing and Prunning* (M-DHP) [4].

2. TINJAUAN PUSTAKA

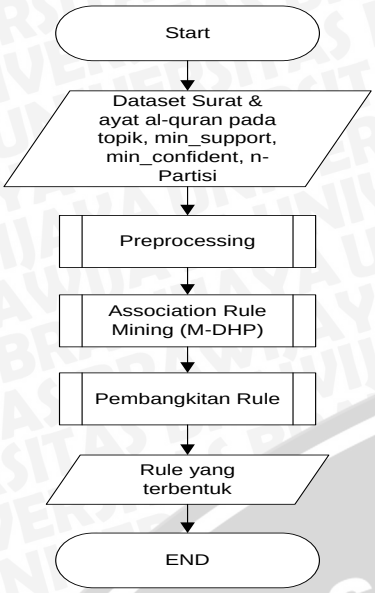
Menurut Kusrini [5] di dalam bukunya "*Algoritma Data Mining*" analisis asosiasi atau *assication rule mining* adalah teknik data mining untuk menemukan aturan asosiatif antara kombinasi item. Algoritma M-DHP *(Multipass Direct Hashing and Prunning)* sendiri merupakan algortima *assication rule mining* yang dikembangkan oleh John D. Holt dan Soon M. Chung. Algoritma M-DHP ini dikembangkan untuk mencoba mengatasi permasalahan *Association Rule Mining* (ARM) pada *text database*. Hal ini dikarenakan menurut John dan Soon [4] di dalam jurnalnya yang berjudul "*Efficient Mining of Association Rules in Text Databases*" bahwa karakteristik dari *text database* sedikit memiliki perbedaan dari data transaksi perdagangan sehingga menyebabkan besarnya ruang memori yang diperlukan untuk menghitung *frequent itemset*.

Pada penelitian sebelumnya yaitu untuk M-DHP yang dikembangkan oleh John D. Holt dan Soon M. Chung, lebih menitik beratkan pada efisiensi untuk mengurangi besarnya ruang memori untuk menghitung *frequent itemset* dan upaya menekan waktu proses, dimana hal ini berpusat pada proses partisi (k-partisi), sedangkan pada penelitian ini mencoba untuk menganalisa bagaimana pengaruh k-partisi terhadap tingkat kekuatan aturan yang dihasilkan oleh sistem.

3. METODE PENELITIAN

Sistem yang akan dibuat merupakan program yang dapat menemukan asosiasi topik dari ayat-ayat surat dalam Al-Qur'an dengan menerapkan metode M-DHP. Dalam sistem ini nantinya akan menghasilkan rule-rule yang merepresentasikan keterkaitan atau asosiasi antar topik berdasarkan dari bentuk-bentuk transaksi pada setiap surat di dalam Al-Qur'an.

Pada awal dari menjalankan sistem ini tentunya memerlukan berbagai inputan antar lain: input data transaksi, *nilai minimum support*, *minimum confidence*, dan jumlah partisi untuk *multipass*. *Dataset* atau data transaksi pada sistem ini terdiri dari TID dan *itemset*, dimana TID merepresentasikan id transaksi yang berupa nomor surat dan *itemset* yeng berupa kumpulan topik-topik pada tiap surat yang terkandung di dalam Al-Qur'an. Alur pencarian asosiasi topik dari ayat-ayat surat pada Al-Qur'an secara umum dapat dilihat langkah-langkahnya pada gambar 1 dibawah ini:



Gambar 1 *Flowchart* sistem secara umum.

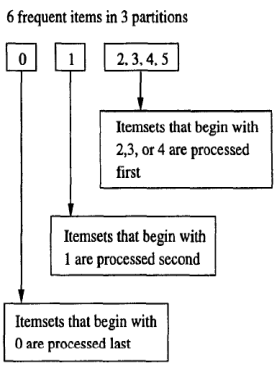
3.1 Preprocessing

Preprocessing dalam system ini meliputi proses pencocokan input dokumen-dokumen topik (*.txt) dengan data topik pada sistem, *tokenizing*, dan Membentuk atau menyusun nomor topik *(Itemset)* dan output dari tahapan *Tokenizing* yang berupa daftar nomor surat *(TID)* menjadi suatu data transaksi

3.2 M-DHP

M-DHP *(Multipass Direct Hashing and Prunning)* sendiri merupakan pengembangan dari DHP *(Direct Hashing and Prunning)*. Berikut ini akan dijelaskan tahapan dari algortima M-DHP yang sesuai dengan gambar 2:

1. Hitung seluruh kombinasi dari masing-masing item di dalam database transaksi dan temukan *frequent 1-itemset* (L_1).
2. Partisi *frequent 1-itemset* ke dalam p partisi $P_1, P_2, \dots, P_p$.
3. Gunakan algoritma DHP untuk menemukan seluruh *frequent itemset* di dalam masing-masing partisi, dimana akan diproses secara urut mulai P_p, $P_{p-1}, \dots, P_1$.



Gambar 2 Partisi k-itemset pada M-DHP

3.3 Pembangkitan Rule

Proses generate rule pada penelitian ini menggunakan *Faster Algorithm*. Menggunakan metode ini karena pada penelitian ini memerlukan metode generate rule yang mampu menghasilkan *consequent* lebih dari satu item. Adapun langkah-langkah dari penerapan metode *faster algorithm* adalah sebagai berikut:

1. Masukan berupa L_k dan k. Nilai dari k disini merupakan nilai k-*itemset*.
2. Dilakukan pengecekan jika k ≥2 maka akan melakukan proses H_1 dan menjalankan prosedur *ap-genrules*.
3. Setelah seluruh *rule* terbentuk, maka tahapan akhir mencari akurasi dari *rule* yang terbentuk

Dimana:
L_k = frequent k itemset
k = k-itemset
H_1 = *consequent* satu *item* dari $L_k x$

3.4 Metode Evaluasi

1. Conviction

Nilai range pada *conviction* ini, berada pada nilai 0.5,...,1,...∞ dengan ketentuan *conviction* dianggap memiliki nilai tak terhingga (*infinite*) apabila nilai dari $conf(A \rightarrow B)$ sama dengan 1. Sama seperti *lift*, apabila *conviction* menghasilkan nilai rule yang menjauh dari 1 maka akan dianggap akurat atau *rule*

tersebut memiliki tingkat kekuatan yang baik.[2]

$$conv(A \rightarrow B) = \frac{1-\sup(B)}{1-conf\ (A \rightarrow B)} \quad (1)$$

Dimana:
A, B : *item*.
Support(B) : nilai *support* dari *item* B.
Conf$(A \rightarrow B)$: nilai *confidence* aturan asosiatif $A \rightarrow B$.

2. Hiper Lift Ratio

Pada *hiper lift* hampir sama seperti *lift*, yaitu apabila nilai *hiper lift ratio* dari suatu rule > 1 maka dianggap akurat atau tingkat kekuatan *rule* yang dihasilkan baik. Adapun mengapa pada *hiper lift* menggunakan nilai δ = 0.99 atau 99%, hal ini bertujuan agar hanya akan menghasilkan rule yang bersifat tidak terpercaya *(independent)* tidak lebih dari 1% dari setiap kasus [4]

$$hyper\ lift\delta(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{\delta\left(support(Y)\right)} = \frac{support\ (x \cup y)}{\delta(support\ (x)\ support\ (y))} \quad (2)$$

Dimana:
δ : 0.99.
Conf$(X \rightarrow Y)$: nilai *confidence* aturan asosiatif $X \rightarrow Y$.
Support(Y) : nilai *support* dari *item* Y.

3. Akurasi

Kualitas dari aturan asosiasi dapat dievaluasi menggunakan rumus akurasi. Akurasi dapat dihitung berdasarkan persentase *error* yang terjadi. Akurasi didefinisikan sebagai berikut :

$$Akurasi = 100\% - Error \quad (3)$$

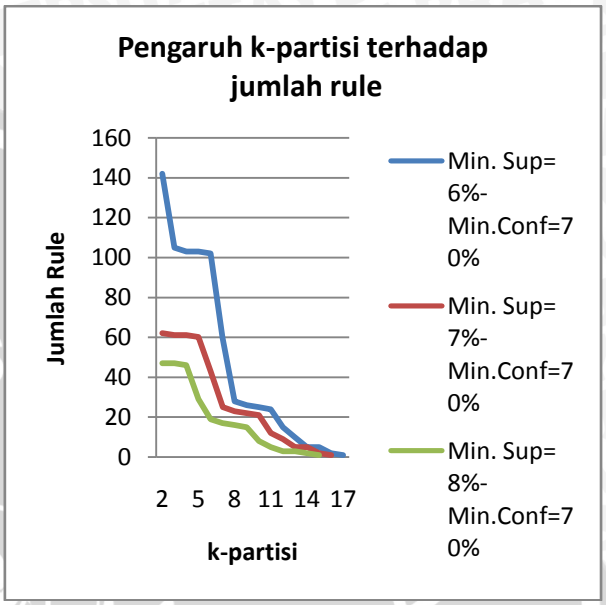
Dimana,

$$Error = (prediksi\ salah / total\ prediksi) * 100\%$$

4. HASIL DAN PEMBAHASAN

4.1 Pengujian Pengaruh k-Partisi Terhadap Tingkat Kekuatan *Rule* dan Jumlah *Rule*.

Pada tahapan ini, hasil pengujian pengaruh k-partisi terhadap jumlah *rule* yang dihasilkan dengan range-partisi 2-partisi hingga n-partisi (tergantung dari batas minimum *support*), variasi minimum *support* 6%, 7% dan 8%, serta minimum *confidence* 70% dapat dilihat pada gambar 3.

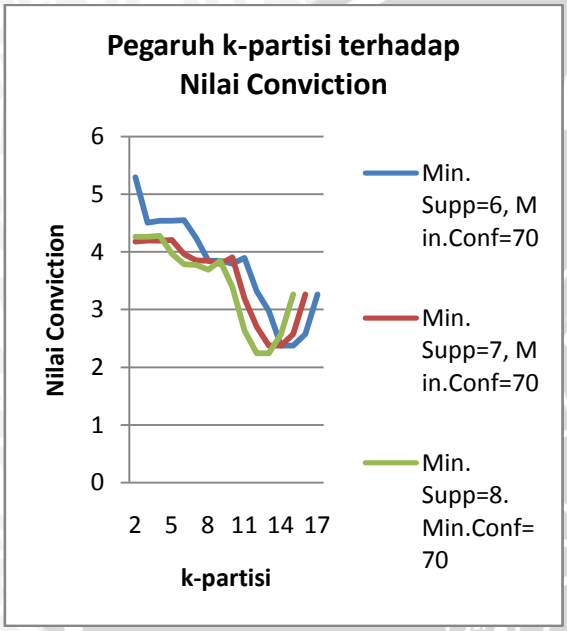


Gambar 3 Pengaruh k-partisi terhadap jumlah *rule*

Sebagaimana telah ditunjukkan pada gambar 3, jumlah *rule* dengan k-partisi=2, minimum *support*=6% dan minimum *confidence*=70% memiliki jumlah *rule* tertinggi, yaitu berjumlah 142 *rule*. Pada grafik minimum *support* 6% dan minimum *confidence* 70% terjadi penurunan jumlah *rule* seiring bertambahnya jumlah k-partisi. Hal ini pula yang terjadi pada minimum *support* 7% dan 8%, yaitu semakin besar k-partisi maka mengakibatkan jumlah *rule* yang dihasilkan juga semakin berkurang.

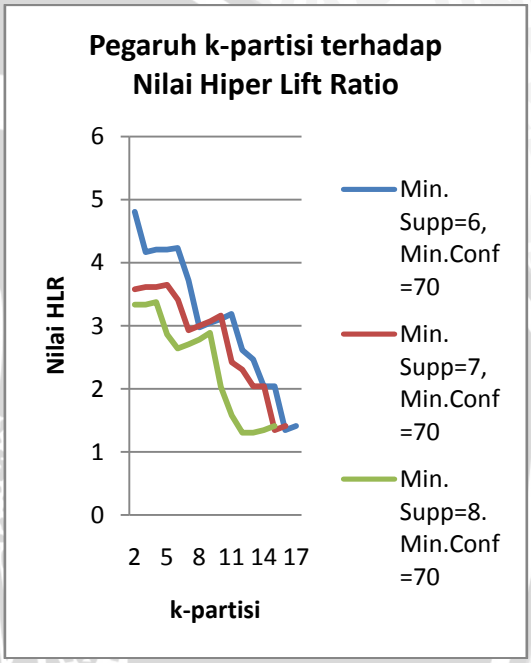
Seperti yang ditunjukkan pada gambar 3, bahwa k-partisi juga berperan dalam hal jumlah *rule* yang dihasilkan pada sistem ini, yaitu semakin tinggi nilai k-partisi maka semakin rendah pula jumlah *rule* yang dihasilkan begitupun sebaliknya, ketika k-partisi rendah maka jumlah *rule* yang dihasilkan juga cenderung meningkat. Hal ini disebabkan pada metode M-DHP terdapat proses reduksi *item* database transaksi dan proses reduksi berkaitan dengan proses partisi karena

pada tahapan partisi memilliki peranan dalam pembentukan *candidate k+1-itemset*. Adapun dalam mereduksi suatu *item* pada database transaksi seperti yang telah dipaparkan pada Bab Metodologi dan Perancangan tergantung dari jumlah dari *item* tersebut, dimana untuk menemukan jumlah tiap *item* database transaksi berdasar pada kombinasi *candidate itemset* yang terbentuk dari *frequent 1-itemset* yang telah terpartisi.



Gambar 4 Pengaruh k-partisi terhadap Nilai *Conviction*

Pada Gambar 4 menunjukkan pengaruh k-partisi terhadap tingkat kekuatan *rule* yang dihasilkan menggunakan *Conviction*, adapun dapat diamati bahwa nilai *conviction* yang dihasilkan untuk mengukur kekuatan suatu *rule* tergolong pada *rule* yang baik atau kuat walaupun terlihat pada gambar 4 terdapat pola-pola grafik yang menurun. Misalkan pada minimum *support* 6% dan minimum *confidence* 70%, untuk k=2 nilai *conviction* berada pada nilai 5,3 dan selanjutnya seiring dengan bertambahnya k-partisi maka grafik juga cenderung menurun namun pada akhirnya meningkat. Berdasar penjelasan tersebut terlihat bahwa terdapat perubahan akibat dari k-partisi, namun tetap tingkat kekuatan *rule* yang telah dihasilkan menggunakan *conviction* pada taraf tingkat kekuatan *rule* yang baik, dimana sesuai dengan yang telah dipaparkan pada Bab Tinjauan Pustaka bahwa nilai range pada *conviction* berada pada nilai 0.5,...,1,...∞ dan jika *conviction* menghasilkan nilai rule yang menjauh dari 1 maka akan dianggap akurat atau *rule* tersebut memiliki tingkat kekuatan yang baik [2].



Gambar 5 Pengaruh k-partisi terhadap Nilai *HLR*

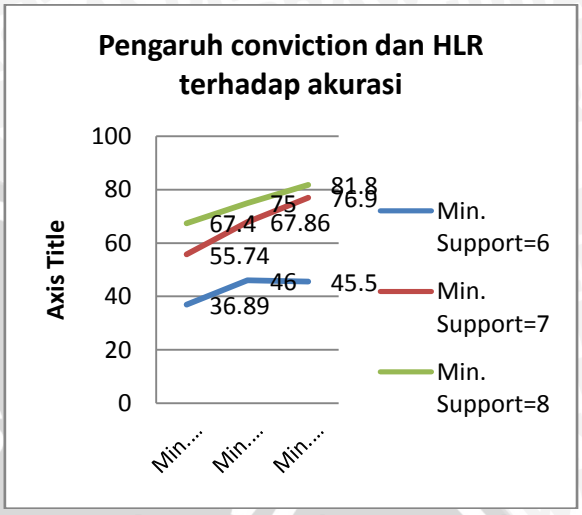
Sebagaimana yang telah ditunjukkan pada Gambar 5 bahwa pengaruh k-partisi terhadap tingkat kekuatan *rule* yang dihasilkan menggunakan *Hiper Lift Ratio*, adapun dapat diamati bahwa nilai *Hiper Lift Ratio* yang dihasilkan untuk mengukur kekuatan suatu *rule* tergolong pada *rule* yang baik atau kuat walaupun terlihat pada gambar 5 terdapat pola-pola grafik yang menurun. Misalkan pada minimum *support* 6% dan minimum *confidence* 70%, untuk k=2 nilai *Hiper Lift Ratio* berada pada nilai 4,8 dan selanjutnya seiring dengan bertambahnya k-partisi maka grafik juga cenderung menurun. Berdasar penjelasan

tersebut, terlihat bahwa terdapat perubahan nilai akibat k-partisi, namun tidak dapat dikatakan bahwa tingkat kekuatan *rule* yang telah dihasilkan menggunakan *Hiper Lift Ratio* cenderung tidak atau kurang kuat karena sesuai dengan yang telah dipaparkan pada Bab Tinjauan Pustaka bahwa apabila nilai *hiper lift ratio* dari suatu rule > 1 maka dianggap akurat atau tingkat kekuatan *rule* yang dihasilkan baik. [4]

Berdasark analisa di atas yaitu pengaruh k-partisi terhadap jumlah *rule* dan pengaruh k-partisi terhadap tingkat kekuatan *rule* yang dihasilkan manggunakan *Conviction* dan *Hiper Lift Ratio*, dapat disimpulkan bahwa perubahan nilai k-partisi memiliki pengaruh terhadap nilai *Conviction* dan *Hiper Lift Ratio* yang dihasilkan, hal ini terlihat dengan adanya pola-pola grafik baik menurun ataupun meningkat. Meskipun terdapat pengaruh k-partisi terhadap tingkat kekuata *tule* yang dihasilkan oleh *Conviction* dan *Hiper Lift Ratio*, terlihat bahwa pada pengujian ini tetap menghasilkan *rule* dengan tingkat kekuatan yang baik, namun dengan nilai k-partisi yang tidak terlalu tinggi juga memiliki keunggulan yaitu semakin banyak kombinasi *rule* yang dihasilkan.

4.2 Pengujian Pengaruh *Conviction*, *Hiper Lift Ratio* dan *Confidence* Terhadap Tingkat Akurasi *Rule*

Pada tahapan ini, merupakan hasil pengujian pengaruh nilai *Conviction* , *Hiper Lift Ratio* dan *Confidence* terhadap akurasi dari *rule-rule* yang dihasilkan. Untuk nilai partisi dipilih pada k=4 agar kombinasi rule yang dihasilkan lebih banyak, sedangkan *rule* yang digunakan pada pengujian ini merupakan *rule-rule* yang nilai *Conviction* dan *Hiper Lift Ratio* memenuhi syarat sebagai *rule* yang memiliki tingkat akurasi yang baik. Untuk variasi nilai minimum *confidence* 70%, 80% dan 90%, sedangkan untuk minimum *support* diambil pada nilai 6%, 7% dan 8%. Hasil pada pengujian ini ditunjukkan pada gambar 6.



Gambar 6 Pengaruh *Conviction* dan *HLR* terhadap Nilai Akurasi

Sebagaimana ditunjukkan pada gambar 6, akurasi pada minimum *support* 8% dan minimum *confidence* 90%, memiliki nilai yang paling tinggi, yaitu 81,8%.

Pada grafik minimum *support* 7% terjadi peningkatan akurasi seiring dengan bertambahnya persentase minimum *confidence*. Untuk minimum *confidence* 70%, akurasi adalah 55,74%, kemudian pada minimum *confidence* 80% dan 90%, akurasi mengalami peningkatan menjadi masing-masing 67,86% dan 76,9%. Meskipun pada minimum *support* 6% terjadi penurunan akurasi yaitu pada minimum *confidence* 80% akurasinya sebesar 46% menjadi 45.5% pada minimum *confidence* 90%, namun penurunan tersebut tidak terlalu signifikan dan cenderung membentuk grafik datar. Oleh karena itu secara umum peningkatan akurasi ini juga terjadi pada grafik minimum *support* 6% dan 8%. Hal ini disebabkan selain *rule-rule* yang digunakan pada pengujian ini merupakan *rule* yang memiliki tingkat kekuatan yang baik berdasarkan pada nilai *conviction* dan *hiper lift ratio* juga karena ketika minimum *support* semakin meningkat maka dalam pembentukan *rule* akan

disaring *itemset* data transaksi yang kemunculannya pada data transaksi tinggi, sedangkan untuk standart minimum *confidence* juga dipilih *itemset* yang memiliki *ratio* kemunculan item *antecedent* dan *consequent* secara bersamaan dengan kemunculan item *antecedent* yang semakin tinggi pula. Sehingga semakin meningkat nilai minimum *support* dan minimum *confidence* maka cenderung akan semakin meningkat pula akurasinya.

5. KESIMPULAN

Berdasarkan hasil penelitian tentang pencarian asosiasi topik dalam ayat Al-Qur'an dengan menerapkan metode *Multipass Direct Hashing and Prunning* (M-DHP), dapat disimpulkan bahwa:

1. Penerapan metode *Multipass Direct Hashing and Prunning* (M-DHP) untuk menemukan asosiasi topik dalam ayat Al-Qur'an meliputi beberapa tahap, yaitu sebagai berikut :
   a. Tahap *preprocessing*, merupakan tahap pembentukan data transaksi yang selanjutnya digunakan pada implementasi penelitian ini.
   b. Tahap implementasi *Multipass Direct Hashing and Prunning* (M-DHP), meliputi prosedur pembentukan kandidat *itemset*, menghitung *support itemset* data transaksi, pencarian *frequent itemset*, proses partisi *frequent 1-itemset* dan proses reduksi database data transaksi.
   c. Tahap *generate rule*, meliputi pembentukan *rule*, prosedur hitung *confidence*, menghitung nilai *conviction* dan *hiper lift ratio* dari tiap *rule.*
2. Nilai k-partisi berpengaruh pada nilai *conviction* dan *hiper lift ratio* dari tiap-tiap *rule*, namun tetap menghasilkan *rule* dengan tingkat kekuatan yang baik. Meskipun hasil dari k-partisi tetap menghasilkan tingkat kekuatan *rule* yang baik, tetapi dalam pemilihan k-partisi sebaiknya tidak terlalu tinggi karena semakin tinggi k-partisi maka kombinasi *rule* yang dihasilkan semakin rendah.
3. Tingkat akurasi tertinggi yang dapat dihasilkan oleh sistem adalah sebesar 81,87% pada minimum *support* 8% dan minimum *confidence* 90%. Selain pengaruh dari *rule* yang digunakan untuk pengujian merupakan *rule* dengan tingkat kekuatan yang baik berdasar pada nilai *conviction* dan *hiper lift ratio* dari setiap *rule*, juga disebabkan oleh semakin meningkatnya minimum *support* dan minimum *confidence* yang mempengaruhi peningkatan akurasi sistem, karena ketika minimum *support* semakin meningkat maka dalam pembentukan *rule* akan disaring *itemset* data transaksi yang kemunculannya pada data transaksi tinggi, sedangkan untuk standart minimum *confidence* juga dipilih *itemset* yang memiliki *ratio* kemunculan item *antecedent* dan *consequent* secara bersamaan dengan kemunculan item *antecedent* yang semakin tinggi pula.

DAFTAR PUSTAKA

[1] Agrawal, S., Srikant, R. 1994. *Fast Algorithm for Mining Association Rules*. Proceedings of the 20th VLDB Conference Santiago, Chile. San Jose.

[2]Azevedo, P.J., Jorge,A.M. 2006.*Comparing Rule Measures for Predictive Association Rules*. Departamento de Informatica Universidade do Porto. Portugal.

[3] Hashler, M., Hornik, K. 2006. *New Probabilistic Interest MeasuresFor Association Rules*. Departement of Statistics and Mathematics WU Vienna University. Vienna.

[4] Holt, J. D., Chung, S. M. 2000. *Efficient Mining of Association Rules in Text Databases*. Department of Computer Science and Engineering Wright StateUniversity .Ohio, USA.

[5] Kusrini, Luthfi, ET. 2009. *Algoritma Data Mining*. Penerbit Andi. Yogyakarta.

[6] Park, J. S., Chen M., Yen, P.S. 1995. *An Effective Hash-Based Algorithm for Mining Association Rules*. IBM Research Center. New York.