ISSN: 2089-1121

ERROR-TOLERANT FASCICLES UNTUK COLLABORATIVE FILTERING

Gunawan, Herman Budianto, Dody Soegiharto, dan Indra Maryati
Jurusan Teknik Informatika
Sekolah Tinggi Teknik Surabaya
gunawan@stts.edu , herman@stts.edu , dody_formelo@yahoo.com , maryati@stts.edu

ABSTRAK

Penelitian ini mempelajari suatu sistem penyusun rekomendasi yang lebih dikenal dengan nama *Collaborative Filtering*. Pembahasan Collaborative Filtering ini akan difokuskan pada penggunaan metode *Error-Tolerant Fascicles* yang terkait erat dengan konsep *association analysis*. Pengimplementasian collaborative filtering dengan metode apapun dapat dibagi menjadi 3 (tiga) tahapan penting, yaitu : *data preparation*, pencarian pola selera user, dan penyusunan rekomendasi. Khusus untuk Collaborative Filtering dengan Error-Tolerant Fascicles, tahap pencarian pola selera user yang dilakukan banyak melibatkan proses association analysis. Data yang digunakan sebagai data *input* untuk pembahasan collaborative filtering ini merupakan dataset selera user terhadap lelucon dan film yang diperoleh dari internet.

Kata kunci : Collaborative Filtering, Error-Tolerant Fascicles, Frequent Pattern Mining, Association Analysis

ABSTRACT

This paper studies a recommender system which is known as collaborative filtering. The Collaborative Filtering study will be focused to Error-Tolerant Fascicles Method which is tightly related with association analysis. Implementation collaborative filtering with any methods can be divided into three main steps, data preparation, user taste pattern search, and recommendation arrangements. Specially for Collaborative Filtering with ETF, finding the user taste patterns step is known as pruning process that involves association analysis concept. The data which are used as data input in this studies are the taste user datasets to jokes and movies which are gained from the Internet.

Keywords : Collaborative Filtering, Error-Tolerant Fascicles, Frequent Pattern Mining, Association Analysis

1. PENDAHULUAN

Collaborative filtering merupakan suatu sistem penyusun rekomendasi yang cukup populer dan berkembang pesat saat ini. Dengan menggunakan collaborative filtering, maka data yang besar, yang telah dikumpulkan dalam waktu yang cukup lama dapat digunakan secara efektif dan bermanfaat. Kumpulan data tersebut dapat digunakan oleh proses collaborative filtering sebagai input untuk penyusunan rekomendasi yang tepat bagi user aktif.

ISSN: 2089-1121

Sebagai suatu sistem penyusun rekomendasi, collaborative filtering tidak jauh berbeda dengan *data mining* dan beberapa metode penggalian pengetahuan yang lain yang menggunakan data sebagai sumber pengetahuan. Perbedaan mencolok collaborative filtering dengan pendekatan-pendekatan tersebut adalah pada collaborative filtering proses pencarian pengetahuan yang pada dasarnya merupakan selera user diperoleh dengan mengkolaborasikan selera dari user-user lain yang hampir sama dengan user aktif.

Error-tolerant fascicles merupakan salah satu dari sekian banyak metode pengimplementasian proses collaborative filtering. Dalam metode error-tolerant fascicles, pendekatan association analysis memegang peranan yang cukup penting. Hampir keseluruhan proses collaborative filtering dengan error-tolerant fascicles didasari oleh konsep dari association analysis.

Metode collaborative filtering dengan error-tolerant fascicles ini didasari oleh satu pemahaman bahwa hampir tidak ada sekumpulan user yang memiliki kesamaan selera 100%. Jadi perlu ada toleransi error yang diberikan terhadap hasil proses collaborative filtering yang digunakan untuk memberikan rekomendasi kepada user aktif.

2. COLLABORATIVE FILTERING DENGAN ERROR-TOLERANT FASCICLES

Teknologi penyusunan rekomendasi kepada user aktif dapat dibagi menjadi 3 (tiga) bagian besar, yaitu: *Information Retrieval*, *Information Filtering*, dan Collaborative Filtering. Ketiga teknologi tersebut berfokus pada bidang dan tugasnya masing-masing. Meskipun demikian, tidak menutup kemungkinan adanya kolaborasi yang saling mendukung dari teknologi-teknologi tersebut.

Sistem collaborative filtering pada dasarnya merupakan sistem penyusun rekomendasi bagi user aktif yang didasarkan pada selera. Jadi pengetahuan yang dicari dalam proses collaborative filtering adalah pola-pola selera dari user-user yang ada. Pola-pola tersebut nantinya akan digunakan proses colaborative filtering untuk menyusun rekomendasi bagi user aktif.

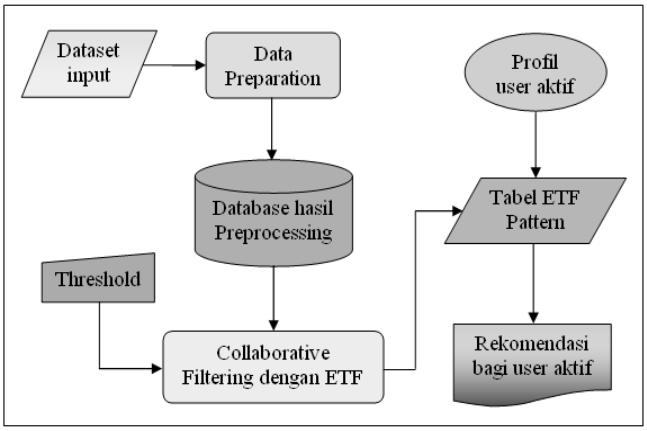
Kekurangan utama yang ada pada collaborative filtering adalah *sparsity* dan rekomendasi user untuk pertama kali. Kedua permasalahan tersebut timbul oleh karena collaborative filtering membutuhkan kolaborasi selera user lainnya dalam menyusun rekomendasi bagi user aktif dengan acuan profil user aktif itu sendiri. Kedua permasalahan tersebut dapat diatasi dengan menggabungkan *information filtering* dengan collaborative filtering.

Pada dasarnya collaborative filtering memiliki hubungan yang cukup erat dengan data mining yang sifatnya prediktif. Banyak dari pendekatan data mining yang juga dapat digunakan untuk implementasi collaborative filtering. Dalam collaborative filtering, pendekatan-pendekatan yang digunakan untuk mengimplementasikannya dibedakan atas 2 (dua) bagian besar, yaitu: algoritma *memory-based* dan algoritma *model-based*.

Metode error-tolerant fascicles merupakan salah satu dari banyak metode dalam algoritma model-based. Metode ini sangat terkait dengan konsep association analysis yang banyak digunakan data mining dalam pencarian frequent pattern mining. Di dalam collaborative filtering, metode error-tolerant fascicles juga tidak jauh berbeda dengan asscociation analysis, yaitu untuk mencari frequent dan *strong ETF*.

ISSN: 2089-1121

Error-tolerant fascicles pada dasarnya merupakan pola selera dari sekelompok user yang terkandung dalam *database* input yang diberikan toleransi error. Sehingga sekelompok user yang menjadi anggota dari suatu fascicles boleh memiliki ketidakcocokan terhadap pola ETF tersebut, maksimal sebesar error yang ditoleransi.



Gambar 1. Collaborative Filtering dengan ETF

Nilai error yang dapat ditoleransi (ε), jumlah item minimum pada tiap fascicles (ϖ_{min}), support minimum fascicles ($fasup_{min}$) dan item ($itemsup_{min}$) merupakan *threshold* yang digunakan dalam collaborative filtering dengan ETF. Keempat threshold itu sebagai batasan agar pola-pola ETF yang dihasilkan oleh collaborative filtering dengan ETF valid, frequent, dan strong.

Dalam collaborative filtering dengan ETF, nilai ε merupakan satu-satunya threshold yang semakin kecil nilainya akan semakin baik bagi rekomendasi yang akan dihasilkan (ε>=1). Hal ini disebabkan karena besarnya nilai ε menyatakan jumlah probabilitas rekomendasi maksimum yang mungkin diberikan kepada user aktif.

Berbeda dengan ε, tiga threshold lainnya akan semakin baik bagi output collaborative filtering dengan ETF jika nilainya relatif besar (besarnya nilai disesuaikan dengan karakteristik database input). $Fasup_{min}$ menentukan frekuensi dari tiap pola, $itemsup_{min}$ menentukan besarnya keterkaitan antar item dalam sebuah pola, dan ϖ_{min} menentukan jumlah minimum item yang harus ada pada tiap pola ETF. Semakin besar nilai ϖ - ε dari suatu pola ETF, maka akan semakin akurat rekomendasi yang diberikan oleh pola tersebut kepada user aktif.

3. PERSIAPAN DATA (DATA PREPARATION)

Proses data preparation merupakan tahap pertama yang harus dilakukan pada sebagian besar proses penggalian pengetahuan dari data, termasuk collaborative filtering. Proses ini digunakan untuk mempersiapkan data input menjadi bentuk data yang sesuai untuk proses penggalian pengetahuan yang akan dilakukan demi alasan keakuratan, efisiensi, dan efektifitas.

Secara umum, proses data preparation dapat dibedakan menjadi 4 (empat) proses penting, yaitu: *data cleaning*, *data integration*, *data transformation*, dan *data selection*. Keempat proses tersebut masing-masing menangani tugas yang berbeda-beda, akan tetapi ada kecenderungan terjadi keterkaitan antara satu proses dengan yang lainnya.

Data cleaning merupakan bagian dari proses data preparation yang menangani 3 (tiga) hal utama dalam kaitannya dengan database yang cukup besar. Tiga tugas utama

ISSN: 2089-1121

dari data cleaning tersebut adalah menangani nilai hilang, data tidak konsisten, dan noise yang terjadi pada data. Proses ini berperan penting dalam menjaga keakuratan dari pengetahuan yang kelak akan dihasilkan.

Data integration merupakan proses dari data preparation yang sangat berperan dalam penanganan proses integrasi data. Ada kalanya data input yang digunakan dalam proses collaborative filtering tidak berasal dari satu sumber, sehingga diperlukan penggabungan sumber-sumber data tersebut. Penggabungan data yang dilakukan tidaklah mudah, karena proses penggabungan data sering kali menimbulkan *redundancy* dan harus ditangani oleh proses ini.

Data transformation merupakan proses merubah/mentransformasikan data input menjadi bentuk dan format yang dibutuhkan oleh proses penggalian pengetahuan seperti collaborative filtering. Pada collaborative filtering dengan ETF, proses data transformasi yang dibutuhkan adalah normalisasi dan atribut *construction*. Proses atribut construction digunakan untuk efisiensi proses pengurutan item, sedangkan normalisasi untuk mempersempit range rating yang digunakan.

Data preparation merupakan proses filtering data-data yang tidak diperlukan dalam proses penggalian pengetahuan. Sehingga proses data selection ini sangat mempengaruhi efisiensi dan efektifitas proses penggalian pengetahuan secara keseluruhan.

Penentuan proses data preparation mana saja yang perlu diimplementasaikan sangat ditentukan oleh kasus yang dihadapi. Untuk collaborative filtering dengan ETF, proses data preparation yang perlu diimplementasikan adalah data transformation, data cleaning (nilai hilang dan tidak konsisten), dan data selection (pemilihan atribut data sebagai informasi tambahan).

Tabel 1. Hasil Uji Coba Transformasi Data

Jumlah Record	Nilai yg Di-generate	Waktu Transformasi	
		Atribut Construction	Normalisasi
1.000.029	10.169	3:59:26	1:06:54
1.810.455	300	1:03:38	2:00:40
1.708.993	300	1:02:09	1:53:17
616.912	295	9:46	43:34

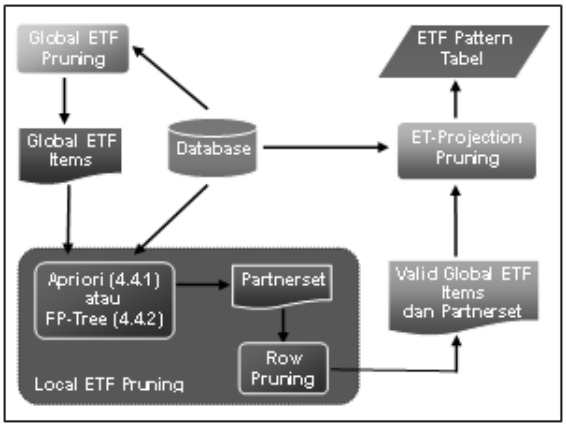
Dari hasil uji coba transformasi data seperti yang terlihat pada tabel 1, maka dapat diambil 2 (dua) kesimpulan penting dalam transformasi data untuk collaborative filtering dengan ETF. Kesimpulan pertama adalah lamanya proses normalisasi yang dibutuhkan sangat ditentukan oleh jumlah *record* yang ada. Kesimpulan kedua adalah lamanya waktu yang dibutuhkan untuk atribut construction tidak hanya ditentukan oleh jumlah nilai yang harus di-*generate*, melainkan juga jumlah record yang ada (hal ini disebabkan karena nilai-nilai yang di-generate tersebut harus di-*update* ke tiap record yang ada).

4. PRUNING COLLABORATIVE FILTERING DENGAN ETF

Proses pruning merupakan tahapan yang pengimplementasiannya memerlukan waktu paling banyak dibandingkan dengan tahapan yang lain. Proses ini bertugas mencari pola-pola ETF yang terkandung dalam database input. Proses pruning dapat

ISSN: 2089-1121

dibedakan menjadi 3 (tiga) tahapan penting, yaitu : (1) *Global ETF* pruning, (2) *Local ETF* Pruning, dan (3) *ET-Projection* Pruning. Masing-masing tahapan tersebut terkait satu sama lain, disebabkan karena hasil dari tahapan sebelumnya akan digunakan sebagai input proses bagi tahapan selanjutnya.



Gambar 2. Pruning Collaborative Filtering ETF

Global ETF Pruning merupakan tahapan pertama dalam pencarian pola ETF pada collaborative filtering dengan ETF. Tahap ini dapat disetarakan dengan pencarian 1-itemset pada proses association analysis. Nilai minimum support yang digunakan sebagai threshold dalam global ETF pruning adalah:

$$Sup(G.ETF\ item) \geq fasup_{min} * itemsup_{min} * |T$$

Dengan nilai threshold tersebut, maka item-item yang dihasilkan oleh global ETF pruning merupakan frequent dan strong item. Hasil dari global ETF pruning merupakan input bagi proses pencarian partnerset pada local ETF pruning. Proses pencarian partnerset dalam local ETF pruning pada dasarnya dapat disetarakan dengan proses pencarian 2-itemset dengan nilai threshold:

$$Sup(A,B) \geq (2 * itemsup_{min} - 1) * fasup_{min} * |T|$$

dimana: $A, B \in G.ETF\ items$

Dari kedua nilai threshold tersebut dapat dilihat bahwa association analysis pada data mining tidak sama persis dengan collaborative filtering dengan ETF, karena collaborative filtering dengan ETF menggunakan threshold yang berbeda dalam pencarian itemset-nya.

Proses dalam local ETF pruning dilanjutkan oleh proses yang dikenal dengan *row* pruning. Proses ini melakukan eliminasi terhadap global ETF item yang memiliki jumlah partner kurang dari ϖ_{min}-ε. Proses eliminasi pada row pruning ini didasari pada pemahaman bahwa setiap pola-pola ETF yang di-generate harus memiliki jumlah item lebih dari ϖ_{min}.

Hasil dari proses row pruning tersebut adalah daftar global ETF item yang baru (L), yang mengeliminasi global ETF item hasil global ETF pruning yang memiliki partner kurang dari ϖ_{min}-ε. Global ETF items dari local ETF pruning inilah yang merupakan global ETF items yang valid, yang akan digunakan untuk mencari pola-pola ETF pada ET-Projection pruning.

Dari hasil uji coba local ETF pruning yang ada pada tabel 2, dapat disimpulkan bahwa besarnya database input sangat mempengaruhi lamanya proses local ETF pruning, baik menggunakan algoritma *apriori* maupun *FP-Tree*. Untuk algoritma

ISSN: 2089-1121

apriori, lamanya proses lebih ditentukan oleh banyaknya global ETF items yang ada. Hal ini berbeda dengan algoritma FP-tree yang harus melakukan duplikasi database input di *memory*, sehingga besarnya database input sangat berpengaruh.

Tabel 2. Hasil Uji Coba Local ETF Pruning

Global ETF items	Jumlah Record relasional	Partnerset	
		Apriori	FP-Tree
2.780	1.000.029	5:06:44	6:12:55
232	1.810.455	3:04:18	16:23:44
242	1.708.993	2:55:46	14:15:44
81	616.912	4:26	5:31:06

Setelah kedua proses tersebut selesai dilakukan, maka langkah terakhir yang perlu diimplementasikan adalah melakukan ET-Projection pruning. Proses ini membutuhkan global ETF items berikut partnersetnya yang dihasilkan oleh local ETF pruning dalam pencarian pola-pola ETF yang ada pada database input.

Pola-pola ETF yang dicari dalam ET-Projection pruning adalah pola-pola ETF yang memenuhi keempat nilai threshold yang telah ditentukan. Pola-pola ETF yang dihasilkan adalah pola-pola ETF yang valid, frequent, dan strong. Setiap pola-pola ETF yang di-generate oleh proses ET-Projection pruning memiliki ketentuan sebagai berikut:

1. Sup(Pola ETF) >= $fasup_{min}$ * |T|
2. ϖ (Pola ETF) >= ϖ_{min}
3. ε (Pola ETF) <= ε
4. Sup(item $\in$ pola ETF) >= $itemsup_{min}$* |F|

Dalam ET-Projection pruning, proses proyeksi tidak dilakukan pada setiap global ETF items yang ada, melainkan cukup hanya (|L|-ϖ_{min}+ε) item pertama (didasarkan atas urutan support) saja. Hal ini pada dasarnya didasari oleh logika bahwa setiap global ETF item sisanya tidak akan memiliki partnerset yang ItemId-nya lebih besar lebih dari ϖ_{min}-ε, sehingga tidak mungkin untuk di-generate sebagai pola ETF.

Semua pola-pola ETF yang dihasilkan oleh ET-Projection pruning ini merupakan pengetahuan yang berharga yang dapat digunakan untuk menyusun rekomendasi bagi user aktif. Validitas dari semua pola-pola ETF yang dihasilkan tersebut sepenuhnya bergantung pada berapa lama selera user-user penggunanya tidak berubah secara signifikan.

5. PENYUSUNAN REKOMENDASI

Proses penyusunan rekomendasi ini merupakan proses terakhir dari collaborative filtering dengan ETF dan langsung berhubungan dengan user. Output rekomendasi yang diberikan kepada user aktif tidak selalu satu, akan tetapi dimungkinkan adanya beberapa alternatif rekomendasi berdasarkan prioritas atau keakuratannya terlebih dahulu.

Tidak semua pola ETF yang dihasilkan oleh proses pruning dapat digunakan untuk menyusun rekomendasi bagi user aktif, akan tetapi hanya pola-pola ETF yang berhubungan dengan profil dari user aktif. Pola-pola ETF tersebut adalah pola-pola ETF yang memiliki error terhadap profil dari user aktif kurang dari ε item. Semakin besar nilai ε akan menambah jumlah probabilitas item rekomendasi setiap pola ETF-nya, sehingga mengurangi tingkat keakuratan rekomendasi.

ISSN: 2089-1121

Dalam collaborative filtering dengan ETF, tidak semua user aktif dapat diberikan rekomendasi. Hal ini disebabkan karena pencarian pola-pola ETF sangat dipengaruhi nilai-nilai threshold yang ada. Sehingga hanya user-user yang profilnya memenuhi threshold-threshold tersebut yang dapat diberikan rekomendasi.

6. PENUTUP

Setelah melakukan beberapa uji coba pada collaborative filtering dengan ETF, ada beberapa kesimpulan penting yang dapat diambil. Beberapa kesimpulan tersebut adalah bahwa proses normalisasi dan atribut construction, serta kompleksitas database input sangat menentukan efisiensi dan efektifitas dari keseluruhan proses collaborative filtering dengan ETF. Di dalam collaborative filtering dengan ETF, keempat nilai threshold yang digunakan sangat mempengaruhi efisiensi proses pruning, dan kualitas dari pola-pola ETF yang dihasilkan maupun rekomendasi yang dapat diberikan kepada user aktif. Semakin besar nilai $fasup_{min}$, $itemsup_{min}$, dan $\varpi_{\min}$ serta semakin kecilnya nilai ε akan membuat keakuratan rekomendasi dan pola-pola ETF semakin baik. Untuk pencarian partnerset dalam collaborative filtering dengan ETF, algoritma apriori kerap kali lebih efisien daripada FP-tree karena tidak perlu melakukan duplikasi database input di memori.

Saran penting yang dapat diberikan yaitu dalam penentuan nilai threshold sebaiknya perlu disesuaikan dengan karakteristik user dalam database input. Hal ini untuk menjaga efisiensi dan agar pengetahuan yang dihasilkan tidak terlalu minim. Selain itu, penentuan metode pencarian partnerset yang tepat dapat mempersingkat waktu proses yang dibutuhkan. Algoritma apriori lebih baik digunakan jika jumlah global ETF items hanya sedikit, akan tetapi jika ukuran database input tidak terlalu besar, algoritma FP-tree akan lebih baik. Oleh karena akses database memakan waktu lebih lama daripada akses memori, maka sebaiknya penggunaan akses database dapat diminimalisir.

DAFTAR PUSTAKA

Han Seng Chee, Sony. *Rec Tree: A Linear Collaborative Filtering Algorithm*. September. 2000. http://www.cs.sfu.ca/CC/470/qyang/lectures/CF%20Ref/chee.pdf

Han, Jiawei dan Kawler, Mixheline. *Data Mining: Concepts and Techniques*. Academis Preas, USA. 2001.

Han, Jiawei, Pei, Jian, dkk. *Mining Frequent Pattern without Candidate Generation: A Frequent-Pattern Tree Approach.* Mei. 2000. http://www.sce.carleton.ca/faculty/ajila/5703/Database_Mining/F-tree-Mning.pdf

H.V. Jagadish, J. Madar dan R. T. Ng. *Semantic Compression and Pattern Extraction with Fascicles*. September. 1999. http://www.acm_org/sigmod/vldb/conf/1999/P16.pdf

Melville, Prem, J.Mooney, Raymond dan Nagarajan, Ramadass. *Content-Boosted Collaborative Filtering for Improved Recommendations*. Juli. 2002. http://www.cs.utexas.edu/users/ml/papers/cbcf-aaai-02.ps.gz

Möhring, Michael. *Data Mining*. Desember. 2002. http://www.uni-koblenz.de/~moeh/lehre/ws0203/dm5.pdf

S. Breese, D. Heckerman dan C. Kadie. *Empirical Analysis of Predictive Algorithms for Collaborative Filtering*. Juli. 1998. http://www.research.microsoft.com/users/breese/algsweb.ps

ISSN: 2089-1121

Wang, Zhaoxia. *Collaborative Filtering Using Error-Tolerant Fascicles*. Maret 2001. http://fas.sfu.ca/pub/cs/theses/2001/ZhaoxiaWangMSc.pdf