PENANGANAN MASALAH COLD START DAN DIVERSITY REKOMENDASI MENGGUNAKAN ITEM-BASED CLUSTERING HYBRID METHOD

The Handling of Cold Start and Recommendation Diversity Problem Using Item-Based Clustering Hybrid Method

Gentra Aditya Putra Ruswanda[1], ZK Abdurrahman Baizal [2], Erliansyah Nasution [3]

[1,2,3]Prodi SI Ilmu Komputasi, Fakultas Informatika Universitas Telkom

[1]**gentra.a@gmail.com**, [2]**baizal@telkomuniversity.ac.id**, [3]**rlinst@yahoo.com**

Abstrak

Pada *recommender system* terdapat dua metode yang sering digunakan yaitu *content-based filtering* dan *collaborative filtering*. Metode-metode ini memiliki kelebihan dan kekurangannya masing-masing. Metode *content-based filtering* memiliki kekurangan dimana hasil rekomendasi yang diberikannya kurang beragam (*diversity*) dikarenakan metode ini hanya mengacu pada konten *item* yang direkomendasikan saja. Pada metode *collaborative filtering* terdapat masalah dimana *item* baru yang masih belum memiliki *rating* tidak dapat direkomendasikan karena data *rating* yang dibutuhkan saat proses *rekomendasi* tidak ada.

Pada penelitian ini akan diuji hipotesis dimana kombinasi dari kedua metode tersebut dapat mengatasi masalah *cold start* dan *diversity* yand dimiliki masing-masing metode. Metode yang digunakan untuk mengkombinasikan kedua metode tersebut adalah *Item-based Clustering Hybrid Method* (ICHM). Pengukuran kemampuan ICHM dalam mengatasi masalah *cold start* akan dihitung menggunakan metrik *Mean Absolute Error* (MAE) sedangkan pengukuran *diversity* dari hasil rekomendasinya akan dihitung menggunakan metrik *Intra-List Similarity* (ILS).

Hasil dari penelitian ini menunjukkan bahwa metode ICHM memiliki kemampuan lebih baik dalam menangani masalah *cold start* dibandingkan dengan *collaborative filtering* dengan nilai MAE 1,4522 dan 3,8103. Hasil dari pengujian *diversity* menunjukkan bahwa ICHM memiliki hasil rekomendasi yang lebih beragam dibandingkan dengan *content-based filtering* dengan nilai ILS -3,7187 dan 34,5709.

Kata Kunci : *recommender system, item-based clustering hybrid method, diversification, cold start*

Abstract

There are two common methods used in recommender system which is content-based filtering and collaborative filtering. These methods have its own advantages and weakness. Content-based filtering method has a weakness where its recommendation results are not diverse enough because of its process only rely on the content of the item itself. In collaborative filtering method, there's a problem where a new item that has not been rated by anyone cannot be recommended at all because the rating data that is needed for the process does not exist.

This research will test a hyphotesis which a combination of these two methods can overcome the cold start and diversity problem from its own methods. The method used here to combine them is Item-based Clustering Hybrid Method (ICHM). A measurement used to measure ICHM performance in overcoming cold start problem is Mean Absolute Error (MAE) while the measurement of recommendation diversity use Intra-List Similarity (ILS) metric.

Results of this research showed that ICHM has a better performance in handling a cold start problem compared to collaborative filtering with an MAE of 1,4522 and 3,8103. Result of the diversity test showed that ICHM has a better recommendation diversity rather than content-based filtering with an ILS of -3,7187 and 34,5709.

***Keyword* :** *recommender system, item-based clustering hybrid method, diversification, cold start*

1. Pendahuluan

Recommender system merupakan bagian dari ilmu *information filtering system* yang mana memiliki fungsi untuk merekomendasikan suatu konten kepada audiensinya [1]. Terdapat dua metode yang umum digunakan dalam membangun *recommender system*, yaitu *content-based filtering* dan *collaborative filtering* [7,8]. Kedua metode ini

memiliki kelebihan dan kekuranganya masing-masing.

Content-based filtering memiliki kekurangan dimana hasil rekomendasi yang diberikan kurang beragam (*diversity*) karena hanya mengacu pada konten daripada *item* tersebut [2]. Sedangkan preferensi seseorang terhadap suatu *item* tidak selalu bergantung pada karakteristik maupun konten *item* tersebut [1]. Penelitian pada *recommender system* saat ini telah melakukan eksperimen pada pengukuran kemampuan *recommender system* selain menggunakan akurasi [12] yaitu *diversity* yang akan dihitung dengan metode *intra-list similarity*.

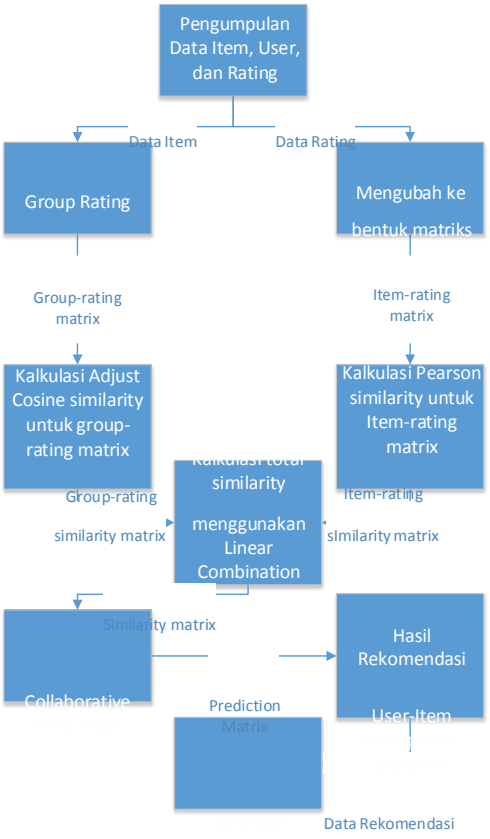
Berbeda dengan *content-based filtering*, metode *collaborative filtering* ini melihat pola kesamaan *rating* yang diberikan tiap *user*. Kekurangan dari metode ini adalah ketidakmampuannya dalam memberikan rekomendasi untuk *item* baru yang belum memiliki data *rating* sama sekali, atau biasa disebut masalah *cold start* pada *item* [2, 9, 10, 11].

Penelitian ini akan menguji hipotesis bahwa kombinasi dari kedua metode tersebut dapat mengatasi masalah *cold start* dan *diversity*. Penelitian ini akan menggunakan metode *hybrid* yang bernama *Item-based Clustering Hybrid Method* (ICHM). Metode ini mengkombinasikan pendekatan *content-based* dan *collaborative filtering*. Metrik *Mean Absolute Error* (MAE) akan digunakan untuk menguji kemampuan ICHM dalam mengatasi masalah *cold start* sedangkan ILS akan digunakan untuk mengukur *diversity* hasil rekomendasinya.

2. *Item-based Clustering Hybrid Method*

Item-Based Clustering Hybrid Method (ICHM) [2] merupakan metode yang membawa konten dari item ke dalam item-based collaborative filtering. ICHM mengintegrasikan data rating item oleh pengguna dengan tingkat kemiripan dari masing-masing item.

Metode ICHM mengatasi kekurangan-kekurangan yang dimiliki oleh pendekatan Collaborative dan Content-based Filtering. Dari sisi collaborative, ICHM mengatasi masalah cold-start pada pendekatan collaborative yang biasa terjadi ketika terdapat item baru. Pendekatan collaborative memiliki kesulitan untuk memberikan rekomendasi item baru karena belum ada pengguna yang memberikan rating pada item tersebut. Metode ICHM dapat memberikan prediksi rekomendasi item baru karena tidak hanya mengambil informasi dari matriks item-rating namun juga dari matriks group-rating [2] yang merupakan matriks kemiripan item berdasarkan cluster-nya. Sedangkan dari sisi content-based, metode ini mengatasi masalah over-specialization yang biasa terdapat pada jenis metode content-based.



Gambar 1 Gambaran metode ICHM

Berikut merupakan proses kerja pada metode ICHM:

1. Melakukan clustering untuk mengelompokkan item-item, kemudian menggunakan hasilnya untuk menciptakan matriks group-rating.
2. Menghitung tingkat kemiripan (similarity). Metode ini menggunakan tiga perhitungan similarity yaitu: Pertama menggunakan adjusted-cosine algorithm untuk menghitung similarity dari matriks group-rating. Kedua, mengkalkulasi similarity dari matriks item-rating menggunakan Pearson correlation similarity. Terakhir, melakukan kombinasi linier dari kedua perhitungan similarity sebelumnya untuk mendapatkan total similarity.
3. Membuat prediksi suatu item dengan melakukan deviasi rata-rata bobot dari rata-rata neighbour.

2.1. *Group Rating*

Group Rating bertujuan untuk mengelompokkan tiap item ke dalam beberapa cluster. Matriks group rating merupakan matriks probabilitas tiap item masuk ke dalam masing-masing cluster. Matriks inilah Siyang menyediakan informasi content-based metode ICHM [2, 5, 8].

Tiap item akan memiliki atribut-atribut data teks seperti contohnya kategori, nama, dan deskripsi. Atribut-atribut item tersebut dihitung nilai TF IDF-nya menggunakan persamaan (2.5) yang kemudian hasilnya akan digunakan untuk proses clustering menggunakan algoritma Adjusted K-Means Clustering.

Algoritma ini diturunkan dari algoritma K-Means Clustering yang dikembangkan dengan mengaplikasikan teori fuzzy set untuk merepresentasikan hubungan probabilitas antara objek dan cluster pada langkah akhir algoritmanya [2, 5, 8]. Berikut merupakan persamaan untuk menentukan group rating dari suatu item.

$$\Pr o(j,k) = 1 - \frac{CS(j,k)}{MaxCS(k)} \quad \textbf{(2.1)}$$

Keterangan:

$\Pr o(j,k)$ =probabilitas objek *j* masuk *cluster k*

$CS(j,k)$ = *counter-similarity* antar dokumen *j* dengan *cluster k*

$MaxCS(k)$ = nilai maksimum *counter-similarity* pada *cluster k*

Dimana nilai dari CS(j,k) merupakan jarak antara dokumen j dengan centroid cluster k. Nilai jarak tersebut dihitung dengan menggunakan persamaan Euclidean distance berikut

$$d(p,q) = \sqrt{\sum_{i}^{n} =_1 (p_i - q_i)^2} \quad \textbf{(2.2)}$$

Keterangan:

$d(p,q)$ = Jarak dokumen *p* terhadap *cluster q*

n = Jumlah dokumen

p_i = Nilai TF IDF *term i* pada dokumen *p*

q_i = Nilai rata-rata TF IDF *term i* pada *cluster q*

2.2. *Pearson Correlation-based Similarity*

Pearson similarity merupakan salah satu metode kalkulasi similarity yang paling banyak digunakan dalam collaborative filtering. Pearson mengukur derajat hubungan linier antar dua variabel yang ada. Persamaan ini akan digunakan untuk menghitung similarity item berdasarkan nilai rating item yang diberikan oleh penggunanya [5,8].

$$sim(k,l) = \frac{\sum_{u=1}^{m} (R_{u,k} - \overline{R}_k)(R_{u,l} - \overline{R}_l)}{\sqrt{^{m} \quad - \quad ^{2}\sqrt{}}} \quad \textbf{(2.3)}$$

Keterangan:

$sim(k,l)$ = *similarity* antar *item k* dan *l*

m = jumlah pengguna yang memberi *rating k* dan *l*

R_k , R_l = rata-rata *rating* dari *item k* dan *l*

$R_{u,k}$, $R_{u,l}$ = *rating* pengguna *u* pada *item k* dan *l*

2.3. *Adjusted Cosine Similarity*

Adjusted cosine similarity merupakan pengembangan dari cosine similarity yang bertujuan untuk mengatasi masalah perbedaan skala rating yang diberikan. Adjusted cosine similarity memiliki perbedaan dimana setiap nilai rating akan dikurangi oleh nilai rata-ratanya. Berikut merupakan persamaan *adjusted cosine similarity* yang akan digunakan untuk menghitung *similarity item* berdasarkan matriks *group-rating* [5,8].

$$sim(k,l) = \frac{\sum_{u=1}^{m} (R_{u,k} - \overline{R}_u)(R_{u,l} - \overline{R}_u)}{\sqrt{\sum_{u=1}^{m} (R_{u,k} - \overline{R}_u)^2} \sqrt{\sum_{u=1}^{m} (R_{u,l} - \overline{R}_u)^2}} \quad \textbf{(2.5)}$$

Keterangan:

$sim(k,l)$ = *similarity* antar *item k* dan *l*

$\underline{m}$ = jumlah *cluster*

R_u = rata-rata nilai probabilitas *cluster*

$R_{u,k}$, $R_{u,l}$ = nilai probabilitas *cluster u* pada *item k* dan *l*

2.4. *Linear Combination Similarity*

Persamaan berikut akan mengkombinasikan nilai similarity antara kedua hasil similarity yang telah dihitung sebelumnya, yaitu similarity item-rating yang berdasarkan nilai rating dan similarity group-rating yang berdasarkan nilai probabilitas item masuk ke suatu cluster [5,8].

$$sim(k,l) = sim(k,l)_{item} \times (1-c) + sim(k,l)_{group} \times c \quad \textbf{(2.11)}$$

Keterangan:

$sim(k,l)$ = *similarity* antar item *k* dan *l*

c = kombinasi koefisien

$sim(k,l)_{item}$ = *similarity* antar item *k* dan *l* hasil matriks *item-rating*

$sim(k,l)_{group}$ = *similarity* item *k* dan *l* hasil matriks *group-rating*

2.5. *Collaborative Prediction*

$$\sum (R_{u,k} - R_k)$$

$u=1$ $\sum_{u=1}^{m}(R_{u,l}-\overline{R}_l)^2$

Prediksi untuk suatu item kemudian dihitung dengan menggunakan deviasi bobot rata-rata dari rata-rata neighbour. Metode ini menggunakan aturan

Top N untuk memilih N terdekat berdasarkan similarity item [5,8]. Pada masalah cold start, nilai $\bar{R}_k$ akan kosong karena item k tidak memiliki rating sama sekali. Maka nilai $\bar{R}_k$ akan diganti dengan $\bar{R}_{neighbor}$ yang merupakan nilai rating rata-rata nearest neighbors dari item k sejumlah N. Nearest neighbors dari item k ini didapat dari matriks group rating [5,8], yaitu item-item selain k yang berada pada cluster yang sama diurutkan dari besar ke kecil berdasarkan nilai probabilitasnya.

$$P_{u,k} = \bar{R}_k + \frac{\sum_{u=1}^{n} (R_{u,i} - R_i) \times sim(k,i)}{\sum_{i=1}^{n} |sim(k,i)|} \quad \textbf{(2.12)}$$

Keterangan:

$P_{u,k}$ = prediksi untuk pengguna *u* pada *item k*

n = total *neighbour* dari *item k*

$R_{u,i}$ = nilai *rating* pengguna *u* pada *item i*

$\bar{R}_k$ = rata-rata *rating item k*

$sim(k,i)$ = *similarity* antara *item k* dengan *neighbour i*

$\bar{R}_i$ = rata-rata *rating* pada *item i*

3. Evaluasi Pengujian

3.1. Metrik Evaluasi

Pengujian ini akan menggunakan data yang didapat dari MovieLens yang berisikan 5844 rating dari 200 user dan 200 film. Pengukuran akurasi ICHM akan menggunakan metric MAE dengan pembagian data rating menjadi data training dan data testing. Tingkat *diversity* dari hasil rekomendasi akan diukur menggunakan ILS.

3.2. Analisis dan Hasil Pengujian

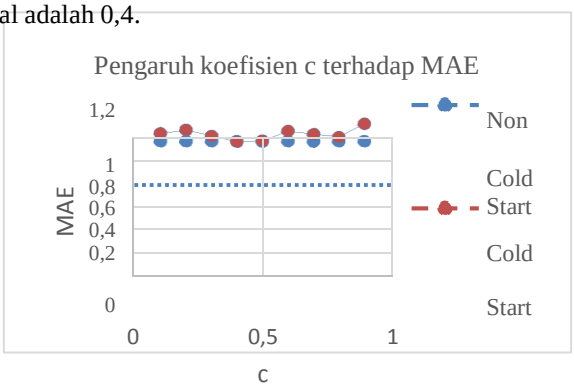
Sebelum memulai pengujian pada masalah cold start dan diversity, perlu didapatkan koefisien kombinasi c yang paling optimal untuk metode ICHM ini.

Maka pengujian awal ini adalah untuk menentukan koefisien c optimal yang akan digunakan oleh metode ICHM ketika akan dibandingkan performanya dengan metode lain. Hal ini dilakukan dengan melakukan iterasi pengujian non cold start terhadap koefisien c mulai dari 0,1 hingga 0,9 dengan selisih c antar iterasi adalah 0,1.

Berdasarkan hasil pengujian yang dapat dilihat pada gambar 3, dapat dilihat bahwa nilai koefisien c tidak begitu mempengaruhi nilai akurasi pada kasus non cold start, namun akan cukup berpengaruh pada kasus cold start. Hal ini dapat terjadi karena pada kasus non cold start data rating item oleh user, yang digunakan untuk membangun item similarity (collaborative), dan data konten item, yang digunakan untuk membangun group similarity (content-based), telah lengkap. Sehingga nilai item similarity dan group similarity sama-sama memiliki nilai dan nilai koefisien kombinasi tidak memiliki pengaruh signifikan.

Sedangkan pada kasus cold start, data rating item oleh user banyak yang kosong menyebabkan nilai item similarity pun menjadi kosong dan tidak dapat merepresentasikan similarity antar item tersebut. Pada kondisi seperti inilah nilai koefisien kombinasi berperan besar menyebabkan nilai akurasi yang dihasilkan lebih bervariasi pada tiap nilai koefisien kombinasinya.

Dari hasil pengujian ini dapat disimpulkan bahwa untuk kasus non cold start nilai koefisien kombinasi yang optimal adalah 0,5. Sedangkan untuk kasus cold start, nilai koefisien kombinasi yang optimal adalah 0,4.



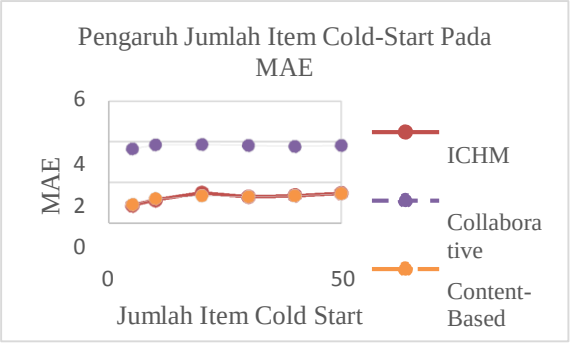
Gambar 3 Pengaruh koefisien c

Setelah didapatkan koefisien kombinasi *c* yang optimal, ICHM dapat dibandingkan dengan kedua metode lainnya untuk dievaluasi akurasinya pada masalah *cold start*.

Masalah cold start adalah kondisi dimana terdapat item baru yang sama sekali belum memiliki rating [2, 9, 10, 11]. Maka dari itu, perlu disimulasikan dataset khusus yang memenuhi kondisi tersebut sebelum memulai pengujian. Hal ini dilakukan dengan cara memilih 5 item dari dataset training kemudian menghapus seluruh data rating yang dimiliki oleh item tersebut dan memindahkannya ke dataset testing.

Berdasarkan hasil pengujian, dapat dilihat bahwa metode ICHM memiliki kemampuan yang lebih baik dibandingkan metode collaborative filtering dalam mengatasi masalah cold start.

Namun ketika dibandingkan dengan metode content-based filtering, ICHM memiliki performa yang sedikit di bawah performa content-based filtering. Dimana rata-rata MAE dari content-based filtering adalah 1,433829, sedangkan ICHM memiliki rata-rata MAE sebesar 1,4522, sedikit lebih besar daripada content-based filtering.



Gambar 4 Pengaruh jumlah *item cold start*

Tabel 1 Hasil kumulatif akurasi seluruh arsitektur

RATA-RATA	MAE	MAPE
ICHM	1,452229	0,402288
Content-based Filtering	1,433829	0,395223
Collaborative Filtering	3,810333	1

Pada pengujian berikutnya, dihitung tingkat diversity dari hasil rekomendasi masing-masing metode menggunakan Intra-List Similarity (ILS). ILS merupakan salah satu metric yang bertujuan untuk mengukur nilai diversity dari himpunan item yang direkomendasikan. Metrik ini menggunakan nilai similarity antar item yang memiliki skala [-1,+1]. Semakin tinggi nilai ILS menunjukkan bahwa himpunan hasil rekomendasi yang diberikan semakin tidak beragam, begitu pula sebaliknya [7].

$$ILS(R) = \frac{1}{2}\sum_{i \in R}\sum_{j \in R} sim(i, j) \qquad \textbf{(3.1)}$$

Setelah dilakukan pengujian, diperoleh hasil nilai ILS dari tiap metode pada tabel 4-1. Berdasarkan tabel tersebut, dapat disimpulkan bahwa metode ICHM memiliki nilai diversity yang jauh lebih baik dibandingkan metode content-based filtering ketika dihitung berdasarkan nilai similarity adjusted cosine. Dalam artian metode ICHM dapat memberikan hasil rekomendasi yang jauh lebih beragam secara konten dibandingkan metode content-based filtering.

Tabel 2 Hasil kumulatif akurasi seluruh arsitektur

ICHM	*Content-based*	*Collaborative*
-3.718763	34.570912	-2.505376

Nilai ILS yang tinggi dari metode content-based filtering ini disebabkan karena metode ini hanya memanfaatkan kesamaan karakteristik konten teks tiap item berdasarkan nilai TF-IDF-nya. Hal ini membuat hasil rekomendasi yang didapat cenderung mirip jika dihitung nilai ILS-nya menggunakan adjusted cosine similarity karena similarity ini juga menggunakan nilai TF-IDF untuk menentukan kemiripannya.

Berbeda dengan metode collaborative filtering yang memanfaatkan kesamaan nilai rating antar item-nya untuk menentukan rekomendasinya. Metode ini menghasilkan rekomendasi yang mirip berdasarkan nilai rating-nya bukan mirip berdasarkan karakteristik kontennya.

ICHM juga mendapatkan nilai ILS yang rendah karena turut memperhitungkan kesamaan rating antar item seperti collaborative filtering. Nilai similarity ini kemudian dikombinasikan dengan adjusted cosine similarity yang melihat kesamaan konten item-nya.

4. Kesimpulan

Dengan mengkombinasikan kelebihan dari kedua metode, ICHM dapat mengatasi masalah *cold start* yang dihadapi oleh metode *collaborative filtering* dan masalah *diversity* yang dihadapi oleh metode *content-based filtering*. Kelebihan dari metode *content-based filtering* berhasil mengatasi masalah *cold start* yang dimiliki oleh *collaborative filtering*. Kelebihan dari metode *collaborative filtering* berhasil mengatasi masalah *diversity* yang dimiliki oleh metode *content-based filtering*.

Namun ketika ICHM dibandingkan dengan kelebihan dari masing-masing metode, didapat perbedaan yang tidak signifikan. Pada masalah *cold start*, metode *content-based filtering* sedikit lebih baik dari ICHM. Pada masalah *diversity*, metode ICHM tidak berbeda jauh dengan metode *collaborative filtering*.

Untuk pengembangan berikutnya, perlu dievaluasi masalah *cold start* pada *user* baru dengan menggunakan metode *User-based Clustering Hybrid Method* (UCHM). Selain itu mencoba untuk mengkombinasikan ICHM dengan *ontology* untuk mendapatkan hasil rekomendasi yang lebih beragam namun tetap relevan dengan *item* sebelumnya.

5. Daftar Pustaka

[1] Jannah, D., Zanker, M., Felfernig, A., Friedrich, G. 2012. *Recommender Systems: An Introduction.* New York: Cambridge University Press.

[2] Li, Q., Kim, B.M. 2012. *An Approach for Combining Content-based and Collaborative Filters.* South Korea: Kumoh National Institute of Technology.

[3] Adomavicius, G., Tuzhilin, A. 2005. *Toward The Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions.* IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, June 2005

[4] Sarwar, B., et al. 2001. *Item-based Collaborative Filtering Recommender System Algorithm*. GroupLens Research Group/Army HPC Research Center, Department of Computer Science and Engineering, University of Minnesota. Miennapolis.

[5] Li, Q., Kim, B.M. 2012. *Clustering Approach for Hybrid Recommender System.* South Korea: Kumoh National Institute of Technology.

[6] Sandoval, Saul. 2012. *Novelty and Diversity Enhancement and Evaluation in Recommender Systems*. Spain: Universidad Autonoma de Madrid.

[7] Ziegler, C., McNee, M., Konstan, J., Lausen, G. 2005. *Improving Recommendation Lists Through Topic Diversification*. Japan: International World Wide Web Conference Comittee.

[8] Kim, B.M., Li, Q., Kim, J.W., Kim, J. 2004. *A New Collaborative Recommender System Addressing Three Problems*. South Korea: Kumoh National Institute of Technology.

[9] Koren, Y., Bell, R., Volinsky, C. 2009. *Matrix Factorization Techniques for Recommender Systems*. U.S.A: IEEE Computer Society.

[10] Zhang, Z.K., Liu, C., Zhang, Y.C., Zhou, T. 2010. *Solving The Cold-Start Problem in Recommender Systems with Social Tags*. Switzerland: Swiss National Science Foundation.

[11] Schein, A.I., Popescul, A., Ungar, L.H., Pennock, D.M. 2002. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM.