Dimensionality Reduction Algorithms for Improving Efficiency of PromoRank: A Comparison Study

Metawat Kavilkrue[1] and Pruet Boonma[2]

[1] Faculty of Engineering, North-Chiang Mai University, Chiang Mai, Thailand
[2] Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
comengi49@gmail.com, pruet@eng.cmu.ac.th

Abstract. It is often desirable to find markets or sale channels where an object, e.g., a product, person or service, can be recommended efficiently. Since the object may not be highly ranked in the global property space, PromoRank algorithm promotes a given object by discovering promotive subspace in which the target is top rank. However, the computation complexity of PromoRank is exponential to the dimension of the space. This paper studies the impact of dimensionality reduction algorithms, such as PCA or FA, in order to reduce the dimension size and, as a consequence, improve the performance of PromoRank. This paper evaluate multiple dimensionality reduction algorithms to obtains the understanding about the relationship between properties of data sets and algorithms such that an appropriate algorithm can be selected for a particular data set. The evaluation results show that dimensionality reduction algorithms can improve the performance of PromoRank while maintain an acceptable ranking accuracy.

1 Introduction

Online marketing becomes an important tool for business and organization [7]. Google[1] and Amazon[2], for instances, relies heavily on online marketing operations, e.g., online advertisement, recommendation and fraud detection. For example, when a user search for a specific term, Google will shows related products in GoogleAds. This can promote the sell of the products because it reflects user's interest [9]. This technique has been used widely not only in business but also public sectors. For instance, when a client borrows a book from a library, the library might want to suggest another related books to them based on their interest. This approach can help boost service satisfiable of clients.

Ranking is a technique to carry out recommendation. It is used widely, for instance, in many bookstores, where top selling books are shown on the front of the stores. This can increase those books selling because people are tend to believe that, because so many other customers already bought these books, they should be good. This is also applied to many other fields in business as well, e.g., American Top Forty, Fortune 500, or NASDAQ-100. Because the number of top ranking is limited, only those who are the best on every dimensions can be in the list. Nevertheless, there are many cases that when consider only a subset of the dimensions, some interesting objects can be found. Consider the following example:

[1] http://www.google.com
[2] http://www.amazon.com

H. Unger et al. (eds.), *Recent Advances in Information and Communication Technology 2015*,
Advances in Intelligent Systems and Computing 361, DOI: 10.1007/978-3-319-19024-2_3

Table 1. Example of multidimensional data

Genre	Year	Object	Score
Science	2012	O_1	0.9
Fiction	2012	O_1	0.2
Fiction	2012	O_2	0.8
Fiction	2011	O_2	0.7
Science	2011	O_2	0.5
Science	2012	O_3	0.4
Fiction	2012	O_3	0.8

Table 2. Target object O_1's subspaces and its ranks

Subspace	Rank	Object Count
{*}	3	3
{Genre=Science}	1	3
{Genre=Fiction}	3	3
{Year=2012}	2	3
{Genre=Science, Year=2012}	1	2
{Genre=Fiction, Year=2012}	3	3

Example 1. (Product Recommendation) It is impossible that Donald Knuth's The Art of Computer Programming series can be in the top list of all books in Amazon store. However, when consider only computer science books with readership toward college students, this book will be ranked on the top list. So, this book series should be recommended only in that particular category and readership.

Example 2. (Online Advertisement) a company wants to promote its product through online advertisement channel, e.g., Google's AdSense. However, the company does not have enough budget to promote the product in all market. However, when consider the sale statistics, the company observe that such product is well accepted by New England market.Therefore, the company can buy advertisement specifically to such market.

From the example, the data space is breakdown into subspaces, e.g., instead of all categories, only New York is considered; hence, the target object, e.g, the salesman, can be the top rank, i.e., top-R, in only some of the subspace. This subspace where the target object is the top rank is called *promotive subspace*.

Table 1 shows a concrete example of a multi-dimensional data set. From the table, there are one object dimension, *Object*, with three target objects, O_1, O_2, O_3. There are two subspace dimensions, *Genre and Year*, and a score dimension, *Score*. Consider O_1 as the target object to promote, Table 2 lists O_1's 6 subspaces and the corresponding rank and object count in each subspace. The rank is derived from the sum-aggregate score of all objects in the subspace. For example, in {Science, 2012}, O_1 ranks 1st

because the score of $O_1 : 0.9 > O_2 : 0.5 > O_3 : 0.4$. Object count is the number of objects in that subspace. Thus {Science, 2012} is a promotive subspace of O_1.

Thus, given a target object, the goal is to find subspace with large promotiveness, i.e., subspace where the target object is top-R. For example, observe that O_1, which is ranked third in all dimensions ({*}), should be promote in subspace {Science} because it is the first rank. In other words, {Science} is a promotive subspace of O_1. The problem of finding large promotiveness subspace is formally defined in section 2.

PromoRank [10] proposes to use subspace ranking for promoting a target object by finding a subspace where the target object is in Top-R ranking. Section 3 discusses PromoRank in detail. However, the computation complexity of PromoRank is exponential to the dimension size, this paper proposes to use dimensionality reduction algorithms, such as principal component analysis (PCA) and factor analysis (FA), to reduce the number of dimensions before performing PromoRank. This approach is explained in Section 4. Dimensionality reduction in recommendation system, in generally, has been studied for many years, e.g., in [1, 5, 8]. In particular, a dimensionality reduction algorithm, i.e., PCA, is used successfully to reduce the subspace ranking used in PromoRank [6]. This paper further investigate this approach by comparing the impact of three well known dimensionality reduction algorithms on the performance of PromoRank on different datasets. The evaluation in Section 5 shows that different algorithms are suitable to different datasets, based on the datasets' properties.

2 Problem Definition

Consider a d-dimensional data set $\mathcal{D}$ with the size of n, each tuple in $\mathcal{D}$ has d **subspace dimension** $\mathcal{A} = \{A_1, A_2, ..., A_d\}$, **object dimension** I_o and **score dimension** I_s. Let $\text{dom}(I_o) = \mathcal{O}$ is the complete set of objects and $\text{dom}(I_s) = \mathbb{R}^+$. Let $S = \{a_1, a_2, ..., a_d\}$, where $a_i \in A_i$ or $a_i = *$ ($*$ refers to any value) is a **subspace** of $\mathcal{A}$. In Table 1, $\mathcal{O} = \{O_1, O_2, O_3\}$, $\mathcal{A} = \{A_{\text{genre}}, A_{\text{year}}\}$. An example of S is {Genre=Science, Year=2012}

As a consequence, S induces a projection of $\mathcal{D}_S \subseteq \mathcal{D}$ and a subspace of object $\mathcal{O}_S \subseteq \mathcal{O}$. For example, when $S = \{Genre = Science, Year = 2012\}$, $\mathcal{O}_S = \{O_1, O_3\}$.

For a d-dimensional data, all subspaces can be group into 2^d cuboids. Thus, S belongs to a d'-dimensional cuboid $\mathcal{A}'$ denoted by $A'_1 A'_2 ... A'_{d'}$, iff S has non-star values in these d' dimensions and star values in the other $d - d'$ dimensions.

Then, for a given **target object** $t_q \in \mathcal{O}$, $\mathcal{S}_q = \{S_q | t_q \in \mathcal{O}_{S_q}\}$ is the set of **target subspace** where t_q occurs. For example, O_1 in Table 2 has 6 target subspaces, as in Table 1, subspace {2011} is not a target subspace because O_1 does not occur in it.

There are many ways to measure the promotiveness of objects in each subspace. One way to measure the promotiveness is percentile-rank, calculated from inverse of the rank of the target object in the subspace times distinct object count, i.e.,

$$P = Rank^{-1} \cdot ObjCount. \tag{1}$$

For example, in Table 2, promotiveness of subspace {Genre=Fiction} is $\frac{1}{3} \cdot 3 = 1$ while the promotiveness of subspace {Genre=Science, Year=2012} is $\frac{1}{1} \cdot 2 = 2$.

Finally, the definition of the promotion query problem is, *given data set $\mathcal{D}$, target object t_q, and promotiveness measure* P*, find the top-R subspaces with the largest promotiveness values*.

This promotion query is a challenging problem because it has a combinatorial nature, the number of combination of subspace with multiple dimensions can increase exponentially. The brute-force approach that enumerates all subspaces and compute the promotiveness in each subspace is prohibitive.

3 PromoRank Algorithm

To address the aforementioned challenge, **promotion analysis through ranking (PromoRank)** [10] utilizes the concept of subspace ranking, i.e., ranking in only selected dimensions. PromoRank consists of two phases: aggregation and partition. Algorithm 1, shows the pseudo-code of PromoRank [10].

Algorithm 1. PromoRank($t_q, S, \mathcal{D}, O_S, d_0$)

```
Input: target object t_q, subspace S, data set D, object set in current
   subspace O_S, current partition dimension d_0
Output: Top-R promotive subspaces Results
1: Results ← ∅
2: if |D| < minsup ∨ t_q ∉ O then
3:   return
4: end if
5: Compute Rank and P
6: Enqueue (S, P) to Results
7: for d' ← d_0 + 1 to d do
8:   Sort D based on d'-th dimension
9:   for all value v in d'-th dimension do
10:     S' ← S ∪ {d' : v}
11:     PromoRank(S', D_S', O_S', d')
12:   end for
13: end for
```

In aggregation phase, if the size of data set is not less than a threshold (*minsup*) and t_q is in the given current subspace, then, promotiveness P of a subspace S is computed and kept in *Results* priority queue. From Algorithm 1, Rank and P of the target object are computed for the input subspace S (Line 5). In particular, Rank can be measured from the rank of the target object in the subspace and P is calculated using Equation 1. Then, S and P are inserted into the priority queue, where P is the key (Line 6). This priority queue maintains the top-R results.

In partition phase, the input data is iteratively processed for an addition dimension (d'). Then, for each distinct value on the d'-th dimension, a new subspace is defined and processed recursively. In particular, the input data $\mathcal{D}$ is sorted according to the d'-dimension (Line 8). Then, $\mathcal{D}$ can be projected into multiple partition, corresponding to the distinct values on the d'-t dimension. A new subspace S' is defined for each partition (Line 10). Then, PromoRank recursively computes over subspace S' (Line 9).

At each recursion, the aggregation phase runs in $O(|\mathcal{D}| + |\mathcal{O}|)$ and the partition phase runs in $O(|\mathcal{D}|)$. Given that there are d dimension, the number of recursion will be 2^d. Thus, the computational complexity of this algorithm is $O(2^d(|\mathcal{D}| + |\mathcal{O}|))$.

4 Dimensionality Reduction for PromoRank

In order to further improve PromoRank, this paper proposes to reduce the number of dimensions (d) of the data set. From the computational complexity of PromoRank, reduce dimensions should impact the performance greatly [2,3]. Moreover, this approach can be performed as a pre-processing for PromoRank; thus, it can be combined with the pruning approaches. However, not all reduction algorithms can be applied to all data set, this paper further investigate this approach by comparing multiple algorithms to find suitable algorithm for a data set with particular parameters.

Given a d-dimensional data set $\mathcal{D}$ with subspace dimension $\mathcal{A}$, a dimensionality reduction algorithm, such as PCA and FA, reduces the number of dimension to d^*, such that $d^* < d$, and a reduced data set $\mathcal{D}^*$ is produced with subspace dimension $\mathcal{A}^*$. Please note that, it does not necessary that $\mathcal{A}^* \subset \mathcal{A}$ because the dimensionality reduction algorithm might generate a new dimension for $\mathcal{A}^*$. In other words, there might exists a subspace $S^* = \{a_1^*, a_2^*, ..., a_{d^*}^*\}$ from $\mathcal{A}^*$ where $a_i^* \notin A_j$ for any $A_j \in \mathcal{A}$.

Thus, the top-R promotive subspace from PromoRank with original data set might differ from the top-R promotive subspace with reduced data set. As a consequence, they cannot be compared directly. In order to handle this, a simple mapping scheme is proposed based on the relationship between the original dimensions and reduced dimensions. Suppose that two original subspace dimensions, A_i and A_j, are reduced to a new subspace dimension A_k^*. Consequently, for a top-R promotive subspace contains $a_k^* \in A_k^*$, it will be compared with a subspace that has $a_i \in A_i$ and/or $a_j \in A_i$; together with the common other subspace dimensions. For example, let's assume that the original dimensions are {City, Country, Year}; then, after a dimensional reduction algorithm is performed on the data set, the new dimensions are {Location, Year} where *Location* is reduced from *City* and *Country*. Thus, if PromoRank considers a subspace {location=Lanna} where *Lanna* is reduced from *Chiang Mai* and *Thailand*, then, the subspace {location=Lanna} will be compared with the subspace {City=Chaing Mai}, {Country=Thailand} and {City=Chiang Mai, Country=Thailand}.

Nevertheless, performing dimensionality reduction algorithms incurs extra computational cost.However, dimensionality reduction algorithms such as PCA and FA have lower computational complexity than PromoRank. For instance, PCA that use Cyclic Jacobi's method has complexity of $O(d^3 + d^2 n)$ [4]. The polynomial complexity of PCA is much lower than the exponential complexity of PromoRank. The experimental results in Section 5.5 confirms that the extra computational cost from a dimensionality reduction algorithm is lower than the performance gain from reducing dimensions.

5 Experimental Evaluation

To investigate the impact of dimensionality reduction algorithm on data sets, an experimental evaluation with four data sets, namely, **Top US private collage**[3], **NBA**[4], **Amazon affiliate income**[5] and **Market analysis**[6], was carried out.

A Java version of PromoRank was developed and tested on a computer with an Intel Core2 Duo 3GHz processor and 4GB of memory. The pruning optimization of PromoRank was disabled in the evaluation to remove the impact on the result.

This evaluation investigates three well-known dimensionality reduction algorithms, namely, principal component analysis (PCA), factor analysis (FA) and linear discriminant analysis (LDA). They reduces the number of variables, i.e., dimensions, by measuring the correlation among them. Then, the variables that highly correlate with the others are removed or combined into new variables. The differences among them are mainly the method that each of them use for correlate data, e.g., LDA does not order correlated dimensions by their variance, as in PCA, but instead focuses on class separability. The evaluation was performed on two parts; first part compares the Top-R promotive subspace of dimensional-reduced data sets and original data sets with PromoRank. For all data set, only top-5 promotive subspaces of each target object are considered. The result of this part is presented in Section 5.1, 5.2, 5.3 and 5.4 for Top US private college, NBA, Market Analysis data set and Amazon affiliate income, respectively. This part compares the performance of each algorithms for reducing dimensions of the data set. The second part, presented in Section 5.5, investigates the performance improvement from the dimensionality reduction.

5.1 Top US Private College Data Set

This data set consists of 100 tuples with 8 subspace dimensions, namely *State, Enrollment, Admission Rate, Admission Ratio, Student/faculty Ratio, 4yrs Graduated Rate, 6yrs Graduated Rate* and *Quality Rank*. This data set contains quantitative data, e.g., *6yrs Graduation Rate*, which cannot be used in PromoRank directly because, from Section 2, a subspace dimension has to be a set. Thus, quantitative data have to be converted to categorical data first. In this paper, the number of categories is set to ten. Each quantitative data will be linearly assigned to a category based on its value.

The result of PCA shows that there are two new principle components, i.e., dimensions, that represents four original dimensions, namely, *Grad Rate* and *Ratio*. *Grad Rate* strongly correlates, i.e., has low variance, with *4yrs Grad Rate* and *6yrs Grad Rate*. *Ratio*, on the other hand, strongly correlates with *Admission Ratio* and *Admission Rate*. For FA, there is a strong collation between *4yrs Graduation Rate* and *6yrs Graduation Rate*, so the former one is removed. Finally, LDA reduces the number of dimensions from ten to six.

[3] `http://mathforum.org/workshops/sum96/data.collections/datalibrary/data.set6.html`
[4] `http://www.basketballreference.com`
[5] `http://wps.prenhall.com/esm_mcclave_statsbe_9/18/4850/1241836.cw/index.html`
[6] `http://www.stata.com`

Table 3. Subspace ranking of Top US private college data set

Target	Original Data Set		FA		PCA		LDA	
Object	Subspaces	Ranks	Subspaces	Ranks	Subspaces	Ranks	Subspaces	Ranks
	{*}	1	{*}	1	{*}	1	{*}	1
CalTech	{State=CA}	1	{State=CA}	1	{State=CA	1	{State=CA}	1
	{4yrs Grad R.=70%}	1	{4yrs Grad R.=70%}	-	{Grad R.=85%}	1	{Grad R.=85%}	1
	{6yrs Grad R.=90%}	1	{6yrs Grad R.=90%}	1	{Grad R.=85%}	1	{Grad R.=85%}	1
	{*}	2	{*}	2	{*}	2	{*}	2
Rice	{Enrollment=2}	1	{Enrollment=2}	1	{Enrollment=2}	1	{Enrollment=2}	1
Uni.	{Enrollment=2,	1	{Enrollment=2,	1	{Enrollment=2,	1	{Enrollment=2,	
	6yrs Grad R.=90%}		6yrs Grad R.=90%}		Grad R.=85%}		Grad R.=85%}	1
	{6yrs Grad R.=90%}	2	{6yrs Grad R.=90%}	2	{Grad R.=85%}	2	{Grad R.=85%}	1⋆
	{*}	3	{*}	3	{*}	3	{*}	3
William	{State=MA}	1	{State=MA}	1	{State=MA}	1	{State=MA}	1
College	{Student/Fac. Rt.=80%}	1	{Student/Fac. Rt.=80%}	1	{Student/Fac. Rt.=80%}	1	{Ratio=40%}	1
	{4yrs Grad R.=90%,	1	{6yrs Grad R.=100%}	1	{Grad R.=95%}	1	{Grad R.=95%}	2 ⋆
	6yrs Grad R.=100%}							

Table 3 compares ranks (marked as *Ranks*) of Top-5 promotive subspaces (marked as *Subspaces*) from the original US private college data set and reduced data sets of three target objects. With Williams College as the target object, when the Top-5 promotive subspace of original data are {*4yrs Graduation Rate=90%, 6yrs Graduation Rate=100%* }. In the PCA reduced data, the comparable subspace is {*Graduation Rate=95%*} . First of all, these two subspaces are compared because *Graduation Rate* is the principle component of *4yrs Graduation Rate* and *6yrs Graduation Rate*. Then, to map these two subspaces, the average of the original data set categories, i.e. 95, is assigned to the reduced data set category. As a consequence, it is possible that , when compared with the other object in O, the rank of the target object in the reduced subspace can be different from the original subspace. The differences are marked by a star symbol. However, this mismatch is infrequently happened. Therefore, the result shows that the ranking of Top-5 promotive subspace is mostly maintained even after the dimensionality reduction is performed on the data. The results show that LDA, which can reduce the number of dimension (from ten to six) the most, maintains an acceptable ranking result compared with the original one. Thus, LDA is the most preferred for this data set.

5.2 NBA Data Set

This data set consists of 4,051 tuples with 12 subspace dimensions, namely *First Name, Last Name, Year, Career Stage, Position, Team, Games, Minutes, Assists, Block, Turnover* and *Coach*. The result from PCA and FA dictates that 6 dimensions, *Game, Minutes, Assists, Block, TurnOver* and *Coach* can be removed. Thus, after dimensional reduction, the reduced data set contains only six subspace dimensions, namely, *First Name, Last Name, Year, Career Stage, Position* and *Team*. On the other hand, LDA cannot be applied to this data set because the classification criterion cannot be met. In other words, LDA cannot distinguish between independent and dependent variables.

Table 4 shows results from NBA data set. After six dimensions are removed, the Top-5 promotive subspaces are hardly change in this data set. In particular, the result of

Table 4. Subspace ranking of NBA data set

Target	Original Data Set		FA		PCA		LDA	
Object	Subspaces	Ranks	Subspaces	Ranks	Subspaces	Ranks	Subspaces	Ranks
	{*}	1	{*}	1	{*}	1		
Kareem	{Pos.=Center}	1	{Pos.=Center}	1	{Pos.=Center}	1		
Abdul-	{League=N}	1	{League=N}	1	{League=N}	1	N/A	
Jabbar	{Team=LA Lakers,	1	{Team=LA Lakers,	1	{Team=LA Lakers,	1		
	Year=1978}		Year=1978}		Year=1978}			
	{*}	2	{*}	2	{*}	2		
Michael	{Pos.=Forward}	1	{Pos.=Forward}	1	{Pos.=Forward}	1		
Jordan	{League=N}	2	{League=N}	2	{League=N}	2	N/A	
	{Team=Utah Jazz}	1	{Team=Utah Jazz}	1	{Team=Utah Jazz}	1		
	{*}	251	{*}	251	{*}	251		
LeBorn	{Car. Stg.=Young,	4	{Car. Stg.=Young,	4	{Car. Stg.=Young,	8⋆		
James	Pos.=Guard}		Pos.=Guard}		Pos.=Guard}		N/A	
	{Car. Stg.=Young}	14	{Car. Stg.=Young}	14	{Car. Stg.=Young}	14		
	{League=N}	233	{League=N}	233	{League=N}	258⋆		

FA does not change at all. This result show that, there are some data set that cannot be improved by LDA. So, FA and PCA are the only available choices for such data set.

5.3 Stock Market Data Set

This data set consists of 5,891 tuples with 23 subspace dimensions, namely *Company Name, Industry Name, Ticket Symbol, SIC Code, Exchange Code, Size Class, Stock Price, Price/Piece, Trading Volume, Market Price, Market Cap, Total Debt, Cash, FYE Date, Current PE, Trailing PE, Firm Value, Enterprise Value, PEG Ratio, PS Ratio, Outstanding, Revenues* and *Payout Ratio*. Similar to the previous data set, this data set is converted to categorical data with ten categories for each subspace dimension.

The result from PCA dictates that a subspace dimension, *Price/Piece* can be removed, and there are two new principle components, namely, *Price* and *Forward PE*. *Price* strongly correlates with *Stock Price* and *Market Price*. *Forward PE*, on the other hand, strongly correlates with *Current PE* and *Trailing PE*. FA, on the other hand, can reduce only one dimension. *Stock Price* and *Market Price* are highly correlated, so the later is removed. Similar to the previous data set, LDA cannot improve this data set.

Table 5 shows results from Stock market data set. From the table, there is only few differences between original and reduced data set, marked by a star in the table. Even though, PCA incurs more changes than FA but they are small and acceptable. On the contrarily, PCA can reduces three dimensions, compared with one of FA, so it should perform more efficient. As a conclusion, in some data set, FA can reduce only a few dimension, so, PCA performs better for such data set.

5.4 Amazon Affiliate Income Data Set

To verify the previous observation, the three algorithms was applied to another data set, namely, Amazon affiliate income data set. This data set has 1,384 tuples and 12 dimensions, namely, *Product Line, ASIN, Seller, Date Shipped, Prices, Real Prices,*

Table 5. Subspace ranking of Stock market data set

Target Object	Original Data Set		FA		PCA		LDA	
	Subspaces	Ranks	Subspaces	Ranks	Subspaces	Ranks	Subspaces	Ranks
Bank of America	{*}	1	{*}	1	{*}	1	N/A	
	{Stock Price=$6, Mkt. Price=$6}	1	{Stock Price=$6}	1	{Price=$6}	1		
	{Size Class=10, FYE=31/12/2010}	1	{Size Class=10, FYE=31/12/2010}	1	{Size Class10, FYE=31/12/2010}	1		
	{Current PE=20, Trailling PE=12},	1	{Current PE=20, Trailling PE=12},	1	{Current PE=20, Trailling PE=12},	1		
Greenshift Corp	{*}	2	{*}	2	{*}	2	N/A	
	{Stock Price=$1, Mkt. Price=$1}	1	{Stock Price=$1}	1	{Price=$1}	1		
	{SIC Code=4953}	1	{SIC Code=4953}	1	{SIC Code=4953}	1		
	{Size Class=8, FSE=31/12/2010}	2	{Size Class=8, FSE=31/12/2010}	2	{Size Class=8, FSE=31/12/2010}	2		
AppTech Corp	{*}	3	{*}	3	{*}	3	N/A	
	{Stock Price=$0, Mkt. Price=$0}	5	{Stock Price=$0}	1★	{Price=$0}	1★		
	{Size Class=4}	1	{Size Class=4}	1	{Size Class=4}	1		
	{Ind. Nm.=Softw., Stock Price=$0}	1	{Ind. Nm.=Softw., Stock Price=$0}	1	{Ind. Nm.=Softw., Stock Price=$0}	2★		

Table 6. Subspace ranking of Amazon affiliate income data set

Target Object	Original Data Set		FA		PCA		LDA	
	Subspaces	Ranks	Subspaces	Ranks	Subspaces	Ranks	Subspaces	Ranks
Mr. Spreads. Excel 2007 Library	{*}	1	{*}	1	{*}	1	N/A	
	{Seller=Amazon}	1	{Seller=Amazon}	1	{Seller=Amazon}	1		
	{Date Shipped=Sep09}	7	{Date Shipped=Sep09}	7	{Date Shipped=Sep09}	7		
	{Date Shipped=Nov09, Items Shipped=1}	4	{Date Shipped=Nov09 Items Shipped=1}	4	{Date Shipped=Nov09 Items Shipped=1}	4		
Excel 2007 Power Prog. With VBA	{*}	2	{*}	2	{*}	2	N/A	
	{Ref. Fee Rt.=6},	3	{Ref. Fee Rt.=6},	3	{Fee=7}	2 ★		
	{Prod. Line=Books, Ref. Fee Rt.=8, Revenue=100}	4	{Prod. Line=Books, Ref. Fee Rt.=8, Revenue=100}	4	{Prod. Line=Books, Ref. Fee Rt.=8, Revenue=100}	1★		
	{Seller=3rd Party}	5	{Seller=3rd Party}	5	{Seller=3rd Party}	5		
Herman Miller Mirra Chair; Fully Loaded; Color; Graphite	{*}	3	{*}	3	{*}	3	N/A	
	{Seller=3rd Party}	1	{Seller=3rd Party}	1	{Seller=3rd Party}	1		
	{Seller=3rd Party, Date Shipped=Feb09, Items Shipped=1, Ref. Fee Rt.=8}	2	{Seller=3rd Party, Date Shipped=Feb09, Items Shipped=1, Ref. Fee Rt.=8}	2	{Seller=3rd Party, Date Shipped=Feb09, Items Shipped=1, Fee=10}	1★		
	{Item Shipped=1, Ref. Fee Rt.=8, Revenue=120}	2	{Item Shipped=1, Ref. Fee Rt.=8, Revenue=120}	2	{Item Shipped=1, Fee=10, Revenue=120}	1★		

Referral Fee Rate, Item Shipped, Revenue, Referral Fees and *URL*, Again, LDA cannot be used to improve this data set because the classification criterion cannot be met. FA can be used with this data set but it can remove only one dimension, i.e., Real Price.

Table 7. Performance comparison

Date Set	Execution Time			
	Original Data Set	with FA	with PCA	with LDA
Top Private College	2 seconds	1 second	1 second	1 second
NBA	20.5 minutes	16 minutes	15.2 minutes	-
Stock Market	124 minutes	108 minutes	99 minutes	-
Amazon Affiliate Income	36.5 minutes	29.3 minutes	21.1 minutes	-

PCA indicates two new component, *Prices* and *Fee*. *Prices* is strongly correlated with *Prices* and *Real Prices* in the original data set while *Fee* is strongly correlated with *Referral Fee Rate* and *Referral Fees*.

Even though, PCA incurs more changes in Top-5 promotive subspace ranking that FA, but the changes are small and acceptable. Therefore, PCA is preferred for this data set because it can reduce more dimensions thatn FA.

5.5 Performance Improvement

This section shows the performance of PromoRank. Table 7 shows the comparison between the execution time of PromoRank on the original data set and the execution time of PCA, FA and LDA to produce reduced data sets plus the execution time of PromoRank on the reduced data sets. The result shows that the dimensionality reduction algorithm can improve performance of PromoRank, for about 25% on a large data set.

6 Conclusion

Dimensionality reduction algorithms are introduced to reduce the dimensions of data set in order to improve the performance of PromoRank algorithm. The results confirm that the dimensionality reduction algorithm can reduce the execution time of PromoRank up to 25% while mostly maintains the ranking result. In particular, when a data set can met the classification criterion of LDA, then LDA is the best choices in terms of the number of reduced dimensions. However, if LDA is not eligible, FA should be evaluated next to see the number of dimensions it can reduce. If it can reduce many, then it is the next best choices. Finally, if FA can reduce only one or two dimensions, PCA should be the best choice because, in general, PCA can reduce more dimensions than FA.

References

1. Ahn, H.J., Kim, J.W.: Feature reduction for product recommendation in internet shopping malls. International Journal of Electronic Business 4(5), 432–444 (2006)
2. Ailon, N., Chazelle, B.: Faster dimension deduction. Commun. ACM 53(2), 97–104 (2010)
3. Fodor, I.: A survey of dimension reduction techniques. Tech. rep., Center for Applied Scientific Computing, Lawrence Livermore National Research Laboratory (2002)
4. Forsythe, G.E., Henrici, P.: The cyclic jacobi method for computing the principal values of a complex matrix. In: Transactions of the American Mathematical Society, pp. 1–23 (1960)

5. Kamishima, T., Akaho, S.: Dimension reduction for supervised ordering. In: Proceedings of the International Conference on Data Mining, pp. 330–339. IEEE Press, Hongkong (2006)
6. Kavilkrue, M., Boonma, P.: Improving efficiency of PromoRank algorithm using dimensionality reduction. In: Nguyen, N.T., Attachoo, B., Trawiński, B., Somboonviwat, K. (eds.) ACIIDS 2014, Part I. LNCS, vol. 8397, pp. 261–270. Springer, Heidelberg (2014)
7. Kotler, P., Keller, K.: Marketing Management. Prentice Hall (2008)
8. Symeonidis, P., Nanopoulos, A., Manolopoulos, Y.: Tag recommendations based on tensor dimensionality reduction. In: Proceedings of the ACM Conference on Recommender Systems, pp. 43–50. ACM, Lausanne (2008)
9. Wang, J., Zhang, Y.: Opportunity model for e-commerce recommendation: Right product; right time. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 303–312. ACM, New York (2013)
10. Wu, T., Xin, D., Mei, Q., Han, J.: Promotion analysis in multi-dimensional space. In: Proceedings of the International Conference on Very Large Databases, pp. 109–120. VLDB Endowment, Lyon (2009)