國立交通大學

多媒體工程研究所

碩士論文

一個新的資料特徵產生方法

應用於手機影片推薦之使用者分群之研究

A Novel Attribute Generation Method

for User Clustering in Recommending Mobile Video

研 究 生：李金龍

指導教授：曾憲雄　教授

中 華 民 國　九 十 九　年 七 月

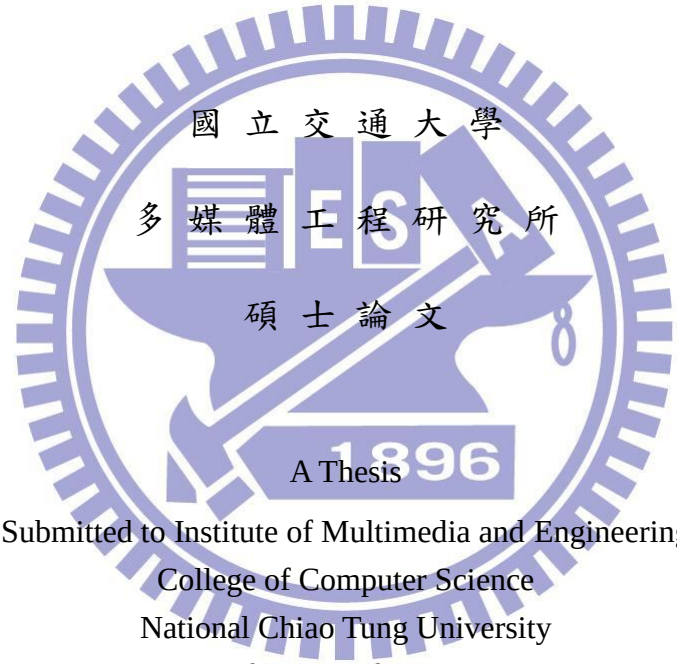
一個新的資料特徵產生方法

應用於手機影片推薦之使用者分群之研究

A Novel Attribute Generation Method for User Clustering in Recommending Mobile Video

研 究 生：李金龍　　Student：Chin-Lung Li

指導教授：曾憲雄　　Advisor：Dr. Shian-Shyong Tseng



國 立 交 通 大 學

多 媒 體 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Multimedia and Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

July 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年七月

一個新的資料特徵產生方法

應用於手機影片之使用者分群之研究

研究生：李金龍　　指導教授：曾憲雄博士

國立交通大學

多媒體工程研究所

摘 要

在現行的推薦系統來說，Amazon 已經證實了透過社群推薦的成功，但是在手機影片推薦系統當中，由於手機影片更新的速度快而且手機的螢幕小的原因使得新的手機影片無法有大量曝光的機會。根據這樣的限制，內容導向過濾偕同合作式過濾推薦方法(Content-based collaborative filtering)就被應用來解決此問題。而原本用來描述影片內容的標籤並不適合用來描述使用者的特徵，這些標籤之間存在著相依或重複性等問題而導致不平衡的使用者分群結果，因此，如何去精煉大眾分類法的標籤成為獨立發散的屬性以用來進行使用者特徵的分群分析就變成很重要的一個議題，在這篇論文當中，一個用於使用者分群的創新的屬性產生方法會透過篩選掉多餘無用的標籤、收縮支配性高的標籤以及歸納隱含性的屬性等方式來凸顯使用者的特徵。而在實驗當中，會透過真實 10906 位手機用戶的 28249 筆交易資料，透過 training data 與 testing data 來驗證。且實驗的結果將證明我們的研究能夠得到更佳的使用者分群，並且能夠提昇使用者族群之間的差異性，以及提昇推薦的命中率。

關鍵字: 推薦系統、社群推薦、分群、屬性產生方法

A Novel Attribute Generation Method

for User Clustering in Recommending Mobile Video

student : Chin-Lung Li Advisors : Dr. Shian-Shyong Tseng

Institute of Multimedia and Engineering

National Chiao Tung University

Abstract

In recent years, the growth of Amazon proved the success of social recommendation. However, in the mobile video recommendation system, the nature of frequent updates of entertainment video and the small screen doesn't allow new contents to have many opportunities for exposure. To solve the issue, the Content-based Collaborative Filtering (CBCF) recommendation approach is applied to solve new item problem. CBCF relies on attributes to characterize users' preferences and makes recommendations according to the log of clustered users with similar interests and the descriptions of contents for users' preferences. Since the common folksonomy-based tag system for video are used to describe the properties of video content but not users' characteristics, the tag dependency problem causes poor user clustering result. Therefore, how to refine the folksonomy-based tags to independently and identically distributed attributes for user characteristics clustering analysis becomes an important issue. In this thesis, a novel attribute generation method based on a taxonomy-based attribute system for user clustering is proposed to reveal the users' characteristics which are screening the redundant tags, shrinking the dominated attributes and generalizing the implicit attributes. In the experiment, the 28249 transactions with 10906 users of the real mobile phone customers have been adopted as training and testing data. The experimental result shows our approach can obtain the representative user characteristics in the clustering results to improve the recommendation.

Keywords: Attribute extraction, Clustering, Recommendation, CBCF, Folksonomy

誌謝

在交大這為期兩年的碩士班研究生涯，可以說是我在求學過程之中，最難忘也最有收穫的一段，在這裡，首先最感謝我的論文指導教授 曾憲雄博士。在曾老師每週超過一次的個別討論的訓練之下，讓我確切的領悟到如何找出問題的核心所在以及正確的研究方法，此外也要特別感謝在口試時給我許多寶貴意見的口試委員：黃國禎教授、楊鎮華教授與彭文志教授，由於口試委員們的意見，才能使得這篇論文更趨完整以及更富有內涵與價值。

再來就是要感謝在這研究的過程當中，帶領我一路剖析問題、尋找方法的學長們：李宗儒學長、翁瑞峰學長、林喚宇學長，感謝你們總是在我卡關不得其解的時候，適時的給予鼓勵以及建議。另外也要感謝其他實驗室的同學們：國彰、杰峰、嘉祥、紹宜、佳榕。在碩士班的這兩年一起苦一起樂一起互相扶持。還有還有其他交大資工系壘的學弟妹們，在假日時能夠一起打球一起宣洩學業上的壓力。感謝我在交大認識的所有人，一起創造了這難忘且無價的回憶！

最後感謝我的家人，在碩士班的日子裡，感謝爸媽從小的栽培，你們從來不給我任何壓力，總是讓我能夠自由的去選擇有興趣想接觸的任何事物，相信經過這兩年的訓練，在未來的日子裡，我都能夠從容面對任何挑戰，永闖難關，不會辜負大家的期待。

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

Due to the tremendous growth of the advanced computer networks and E-commerce services, recommendation systems for increasing trading Volume are becoming critical. Many recommendation systems are successfully applied on the Internet, including recommending books at Amazon.com [7], [12], [9]; movies at Movie Lens [2]; news at VERSIFI [6]; CDs at CDNOW [9]. These successful recommendation systems are built based on Collaborative Filtering approach (CF). The major assumption of CF approaches is that the users will be interested in items which people with similar tastes and preferences liked in the past [1]. However, incapable of recommending new products will decrease the benefits of CF approaches, while large amount of new items continue to update frequently, especially in the on-line video entertainment applications.

The Content-Based Collaborative Filtering (CBCF) recommendation approach is applied to solve the new item problem. CBCF is based on traditional collaborative techniques and also maintains the content-based profiles for each user. These content-based profiles can be used to cluster users and calculate the similarity of products and users' preferences for recommendations. Folksonomy-based tags are common descriptions of video contents and usually used to build the user preference profile. Nevertheless, folksonomy-based tag system which lacks semantic consistency control usually contains the attribute dependency problem. The attributes with dependency can't be used for the user characteristics clustering analysis.

Besides, directly utilizing tag system as the attributes for characterizing users' preferences is not appropriate since the tags for entertainment video content description usually have high dimensionality problem and thus are inappropriate to describe the users' preferences. Refining the folksonomy-based tags aims to solve the

curse of dimensionality which means that data points tend to be more identical as the dimension increases. The high dimensionality of the folksonomy-based tags will cause the bad clustering result. Feature selection techniques in which the most informative dimensions are selected by eliminating irrelevant and redundant ones are usually utilized to reduce the high dimensionality problem. The feature extraction techniques, for more precisely describing the characteristics of each cluster such as projected clustering algorithm [4] [5] and subspace clustering algorithm [12], have been proposed. However, these techniques can only handle the independent attributes from the view of whole feature set and some techniques can only support the post-processing of specific algorithms. Therefore, how to refine the folksonomy-based tags to independently and identically distributed attributes for user characteristics clustering analysis becomes a challenging and important issue. Accordingly, how to design a tag-preference attribute mapping function to generate new user attributes as shown in Figure 1 is defined as the attributes generation problem.

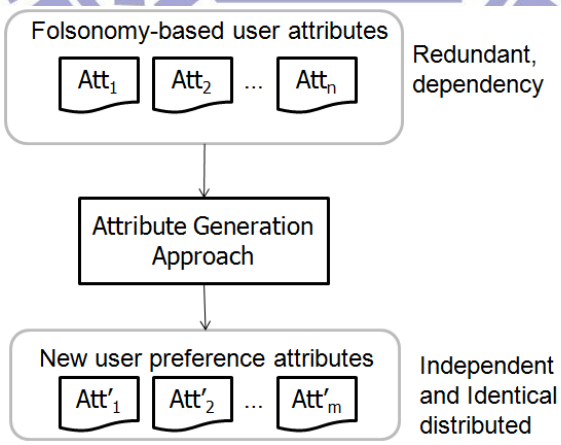


Figure 1. The attribute generation problem

To generate the users' preference attributes, human emotion types have been studied for a long time and our idea of transformation from folksonomy-based content tags to taxonomy-based emotion attributes advised by experts can be used to generate

better attributes for clustering analysis. In this thesis, a novel attribute generation method based on tag-emotion transformation from expert for precisely describing the characteristics of each user and video content from the view of human emotions is proposed. This method consists of three phases. The first phase performs attribute selection by screening the low-informative and high-dependent tags; the goal of the second phase is to shrink the dominated tags, and the third phase transforms the folksonomy-based content tags to taxonomy-based emotion attributes for users' characteristics clustering analysis.

In the experiment, the 28249 transactions with 10906 users of the real mobile phone customers have been adopted as training and testing data. The experimental result shows our approach can obtain the representative user characteristics in the clustering results to fine improve the recommendation system.

The remainder of this thesis is organized as follows: In Chapter 2, we provide a brief overview of recent feature selection techniques and recommendation systems and discuss their strengths and weakness. Chapter 3 describes the problem definition and motivations for solution. Our proposed attribute generation algorithm is proposed in Chapter 4. Chapter 5 presents experiments and performance results in entrainment video recommendation. Our conclusion is given in Chapter 6.

Chapter 2 Related Works

2.1 Collaborative filtering recommendation

The main idea of collaborative filtering (CF) recommendation [16] [11] [25] [10] is that the similar users will like similar products. The CF approaches try to predict the utility of items for a particular user based on the items previously rated by other users. CF analyzed users' behaviors and further provided targeted user with the recommendation according to her/his favorite and relationship among others. Amazon.com [7] developed a CF recommendation system for its bookstore website, where the similar contents were recommended for each content based on the heuristic: similar contents should have similar buyers.

However, although CF approach could be effective in many cases but still has some drawbacks. There are three main problems in the CF approach. Sparsity problem means that the number of already rating items is very small compared to the number of whole items. This phenomenon will lead other items will not be recommended to users. New user problem means recommendation systems recommend items to a person who has no rating experience in the past. Because the user has no rating record, recommendation system can't figure out the characteristic of this user, and system can't identify which cluster the user should belong to. Finally, system would not be able to recommend products similar with the user's characteristics. For the new item problem, if there is no previous users' purchase history on this item, the recommendation system would not be able to recommend it.

2.2 Content-based filtering recommendation

In content-based filtering (CBF) [17] [3] [13] recommendation methods, CBF recommends items by means of the contents' features and users' preferences identified by their historical chosen items and the current clicked item. The CBF approach to recommendation has its roots in information retrieval [20] and information filtering [18] research. Because the significant and early advancements have been made by the information retrieval and filtering communities, many current content-based systems focused on recommending items containing textual information. There are many studies[24] [8] [22] extracted keywords using the retrieval mechanism, e.g., TF/IDF, and recommended contents with similar keywords to those of users' previous features, which can improve the recommendation effectiveness by taking the relations of contents features into account.

One of the measurements for specifying keyword weights in Information Retrieval is the term frequency/inverse document frequency (TF/IDF). TF-IDF weight is a statistical measure in evaluating how important the word is in a document. The importance of a keyword increases when the ratio of its frequency of occurrence in the document to that in the corpus.

The content-based techniques are limited by the features that are explicitly associated with the objects that these systems recommend. Another problem is overspecialization problem; in other words, the system can only recommend items that score highly against a user's profile; the user is limited to be recommended items that are similar to those already rated.

2.3 Hybrid approach

The hybrid approach [26] [13] integrated with CBF and CF concepts becomes the popular mechanism to overcome the problem of pure CF and pure CBF. There are several ways to combine collaborative filtering and content-based filtering methods as hybrid methods.

Content-based collaborative filtering method [1] [14] is one of the most popular hybrid methods in collaboration via content. The main idea of content-based collaborative filtering recommendation is based on traditional collaborative techniques also maintain the content-based profiles for each user. It's a kind of collaborative technique combining content-based characteristic. CBCF approaches follow the CF same principle like CF approach: similar contents should have similar buyers. In CBCF approaches, the users' characteristics modeled by product attribute. So CBCF approaches can overcome the specialization problem of CBF and sparsity problem of CF.

Traditional CBCF[14][15][19] recommendation methods use the tag of contents to model users' profile, and then recommend products according to the clustering result by the attribute of users' profile. So the recommendation result is relevant to the user clustering result.

2.4 Feature selection

Feature selection, which selects an appropriate subset of original feature set plays an important role in the Data mining and Machine learning fields. Feature selection is also widely used in supervised learning and unsupervised learning. However, the unsupervised feature selection is relatively difficult. In the unsupervised configuration, denoising is still a major challenge.

The unsupervised feature selection algorithm for denoising can be categorized two frameworks: wrapper and filter [16]. The wrapper framework uses the clustering method to evaluate the quality of feature and obtain the implicit class information. Information Gain (IG) is one of the criterions of wrapper framework. The information gain of a term measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document.

Although feature selection (FS) for clustering is difficult due to the absence of class labels but FS may lead to more economical clustering algorithms. FS is particularly relevant for data sets with large numbers of features in some applications, such as molecular biology[11] and text clustering applications[25]. Also other approaches as Bayesian approaches for multinomial mixture were proposed in [21] and [26]. A genetic algorithm was used in [17] for FS in K-means clustering.

To solve the new item problem and overspecialization problem and to achieve the idea of social recommendation, CBCF recommendation method will be applied. Nevertheless, how to categorize users with similar characteristics is the most important issue in this thesis. According to the observation of folksonomy-based tag of content and user log in the database, in order to reveal users characteristics we propose an attribute generation process to solve the attribute dependency problem in the folksonomy-based tag system.

Chapter 3 Attribute generation problem

The goal in this thesis is to build an entertainment video recommendation system on the mobile platform, and the principle will follow the observation that people usually trust the recommendations from like-mind friends. Based on the principle of collaborative recommendation, how to find users who had the similar characteristics is important. But the original folksonomy-based tags are used to describe the video contents but are not appropriate to build the user preference profile. In this chapter, the attribute generation problem and our idea of generating item set as attributes of users' characteristic are proposed as followings.

3.1 Problem definition

Assume that there are users U and videos V annotated by tags T, the users' characteristics, the user preference matrix, can be obtained from the user purchasing matrix and the video tagging matrix as following definitions.

Definition 1. The user purchasing matrix

The user purchasing matrix, $P_{|U|\times|V|}$, records the purchase information where U is the set of users and V is the set of videos. In particularly, $P_{i,j} = 1$ denotes ith user purchases jth video; otherwise, $P_{i,j} = 0$.

Definition 2. The video tagging matrix

The video tagging matrix, $A_{|V|\times|T|}$, stands for the relations between videos and tags where V is the set of users and T is the set of tags. In particularly, $A_{i,j} = 1$ denotes ith video is labeled by jth tag; otherwise, $A_{i,j} = 0$.

Accordingly, the users' purchasing frequency can be obtained by the matrix production of $P_{|U|\times|V|}$ and $A_{|V|\times|T|}$.

Definition 3. The user preference matrix

The user tag access matrix, $M_{|U|\times|T|}$, represents the occurrence of tags in user purchase history where U is the set of users and T is the set of tags. In particularly, $M = P \times A$.

In Table 1, the example of partial user log matrix for a mobile video order record is presented. Each row is a transaction of user purchasing a video tagged by certain tags. We aim to extract and generate another attribute set to describe users' characteristic to satisfy the independent attribute requirement for users clustering. However, the main problem is that original tags are proposed for video content annotations which are inappropriate to model users' characteristic. Some issues can be observed in the Table 1. For example, the tag "美聲型(canto)" is rarely used to characterize videos; and "華語(Chinese)" is overused for describing the characterizations of videos. The tags "慢歌(slow songs)" and "抒情(lyric)" often co-occur in the same transaction.

As mentioned above, the attributes of folksonomy-based tag system usually consist of dependency and thus the attributes can't support the user characteristic clustering analysis. Different from the traditional attribute selection issue, tags are used to describe the content and need to be transformed into attributes of users' characteristics.

Table 1. A partial original user log

	慢歌 (slow songs)	抒情 (lyric)	台語 (Taiwanese)	單人 (single)	男 (male)	美聲型 (canto)	華語 (Chinese)	快歌 (fast songs)	偶像型 (idol)	粗礦 (crude)	美麗 (beautiful)	火辣 (hot)	可愛 (cute)	搞笑 (funny)	新秀 (rookie)
U1	1	1	1	1	0	0	1	0	0	1	1	0	0	0	0
U2	1	1	1	0	0	0	1	1	0	1	0	1	0	0	0
U3	1	1	0	0	0	0	1	1	0	1	1	1	0	0	0
U2	0	0	0	0	0	1	1	0	0	1	0	0	0	1	1
U1	1	1	0	1	1	0	1	0	0	1	1	0	1	0	0
U4	1	1	1	1	0	0	1	0	0	1	0	0	0	0	0
U4	1	1	1	0	0	0	1	1	0	0	0	0	1	0	0
U5	0	0	0	1	0	0	1	0	0	1	0	0	0	1	1
U6	1	1	1	0	1	0	1	1	0	1	0	1	0	0	0
U5	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1

The tags with high frequency or low frequency of utilization in every transaction are low-informative tags for discrimination. The co-occurrence tags are the strong group tags which have high dependency and others as weak group tags. Accordingly, the attribute generation problem is defined as followings.

Definition 4. The user preference attribute generation problem

Given the users' transaction data $P_{|U|\times|V|}$ and $A_{|V|\times|T|}$, find a tag-preference attribute mapping function $X_{|T|\times|N|}$ to generate a set of |N| new independently and identically distributed attributes which can transform user preference matrix $M_{|U|\times|T|}$ to user emotion matrix $M'_{|U|\times|N|} = P_{|U|\times|V|} \times A_{|V|\times|T|} \times X_{|T|\times|N|}$ to reveal the required users characteristics.

3.2 The attributes generation approach

To resolve the redundant and dependency attributes, the strategy of cascading processes with redundant attributes cleaning, dominate attributes shrinking, and independent attributes transformation are proposed as followings

(1) Redundant tag screening process

According to the Table 1, it is easy to find the tags: “美聲型(canto)”with low frequency of utilization and “華語(Chinese)” with high frequency of utilization in every transactions. The two tags are redundant for characterizing users. We can determine the tag redundant by evaluating the support of the tag.

According to above observation, we can evaluate the support of each tag to determine whether the tag is useless or not. If a tag’s support value is less than the minimum threshold or more than the maximum threshold, then the tag is useless for characteristic analysis. Therefore, we screen the redundant tags at the beginning of the attribute generation process.

(2) Tag shrinking process

We can find the tags: “慢歌(slow songs)” and “抒情(lyric)” are often co-occurrence in the same transaction. There are many strong group tags will be dominated the characteristic representation. In order to decrease the effect of strong group tags, we propose a tag shrinking process to solve this problem.

Some tags may have similar semantic meaning, so they are always co-occurrence at the same time or in the same content. These tags will dominate other tags. To solve this problem, we can find the shrinking pattern by closed itemset from the original tags then shrink the strong group tags and compensate the weight of these tags.

(3) Tag attribute transformation process

After screening redundant tags and shrinking strong group tags, there are still too many tags may represent the implicit characteristics. As shown in the Table 1, the tags: ”美麗(beautiful)”, ”火辣(hot)”, “可愛(cute)”, “搞笑(funny)”, “新秀(rookie)” may represent some implicit characteristics.

We can define independent emotional attributes from the ICRA which define the content independent attribute by experts. Therefore we could generalize the tags and increase the description power of the other implicit characteristics tags.

3.3 System Architecture

The system architecture of the tag attribute generation is shown in Figure 2. The user-tag raw data and original tags are the input data for the three processes. The tag screening process detects the low information attributes and screens the redundant attributes. Next, the tag shrinking process reduces the impacts of dominated attributes with high dependency. Next, the tag semantic is applied to transform the attributes to another independently and identically distributed attributes. Finally, the new user preference attributes are used in clustering analysis and then applied for the video recommendation.

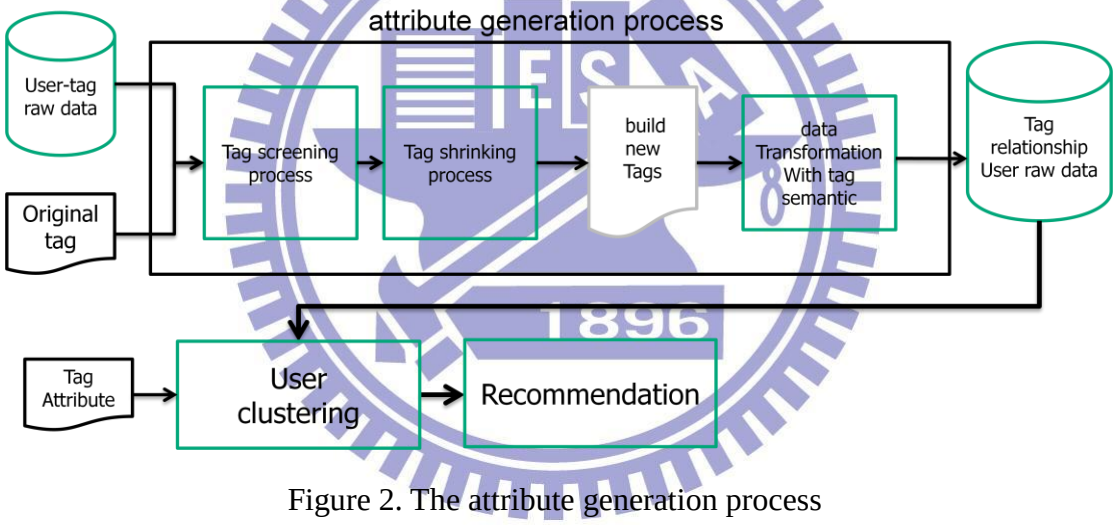


Figure 2. The attribute generation process

(1) Tag screening process

By setting the minimum screening support and maximum screening support, we can filter the redundant and noised tags from the original tags at the beginning of attribute extraction.

(2) Tag shrinking process

Using the screened tag sets from tag screening process, we using closed itemset to find the dominated tags patterns. Then we can build a new tag attribute set.

(3) Data transformation with tag semantic

Define the independent attribute for describing shrunk tags and using the attribute to shrink the un-shrink and spars tags.

(4) User clustering

We apply K-means clustering algorithm using different attributes.

(5) Recommendation

We implement a CBCF recommendation using different user clustering result.

Chapter 4 Methodology

In this chapter, we propose our attribute generation algorithm for meeting the i.i.d. assumption of clustering algorithm and the explainable need of clustering results. This algorithm consists of three phase, including screening low-informative tags, shrinking the dominated tags and transforming folksonomy-based tags into taxonomoy-based human emotion attributes. The corresponding K-means algorithms with different similarity measures are also presented in this chapter. After introducing the notations and definitions used in this thesis, we will present procedures and examples of three phases in order. Finally, attribute generation algorithm will be proposed in the end of this chapter.

4.1 Notations and Definitions

For representing users' preferences by tags, the user tag access matrix is introduced in Definition 3. The influence measure of users on each subset of tags is presented in Definition 4. Definition 5 indicates

Definition 5. Support

Support count is the frequency of occurrence of an itemset $A \subseteq T$, denoted by $\sigma(A)$, Support is fraction of transactions that contain an itemset A, denoted by $S(A) = \frac{\sigma(A)}{\sum_{i,j} P_{i,j}}$. Minimum sup is a user-specified threshold.

Definition 6. Confidence

Confidence of an itemset A over an itemset B, C(A|B), is the ratio of co-occurrences of A and B while B occurs in. In particularly, $C(A|B) = \frac{S(A,B)}{S(B)}$.

Definition 7. Large itemset

A large itemset L is a set of items whose support over a database is larger than minimum support. If $|L| = k$, then L is a large $k-itemset$.

4.2 Three phases of attribute generation algorithm

4.2.1 Tag screening phase

Low-informative tag means that this tag provides no or little information for distinguishing different users' preference. Therefore, it is not beneficial for clustering result and usually the main reason leading to the curse of dimensionality when the amount of low-informative tags is large. Tag screening phase aims to eliminate this kind of tags. The support of a tag measures that the ratio of the occurrences of tag to all transactions. A tag with too low support value (extremely zero) indicates that this tag rarely occurs in purchased videos and hence difficult for describing the differences between users' preferences. Conversely, the similar reason will lead to the low information of tags with too high support. In this phase, to eliminate tags with extreme support value (the maximum and minimum support threshold specified by user) is the core. Example 1 illustrates the procedure of tag screening phase.

Example1：Tag screening process

The user tag access matrix and tag set in Table 2 and Table 3, respectively, are extracted from a customer's purchasing data in a real entertainment platform. This small example consists of 5 customers and 15 tags. The support value of each tag is shown in Table 4.

Here, we assume that the tags occurring in more than 90 percent of videos and less than 10 percent of videos cannot provide sufficient information of distinguishing users' preferences. In this case, the maximum support threshold is set as 0.9 and the minimum support threshold is set as 0.1. Low-informative tags are screened from the original tag set if its support is greater than 0.9 or less than 0.1. Table 5 shows the screened tag set and the corresponding user purchasing matrix is illustrated in Table 6.

Table 2. The user tag access matrix of mobile video purchasing

	慢歌 (slow song)	抒情 (lyric)	台語 (Taiwanese)	單人 (single)	男 (male)	美聲型 (canto)	華語 (Chinese)	快歌 (fast song)	流行樂 (pop)	粗礦 (crude)	美麗 (beautiful)	火辣 (hot)	可愛 (cute)	搞笑 (funny)	新秀 (rookie)
U1	1	1	1	1	0	0	1	0	0	0	1	0	0	0	0
U2	1	1	1	0	1	0	1	1	0	1	0	1	0	0	0
U3	1	1	0	0	1	0	1	0	1	1	1	1	0	0	0
U2	0	0	0	1	0	1	1	0	1	0	0	0	0	1	1
U1	1	1	0	1	1	0	1	0	0	1	1	0	1	0	0
U4	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0
U4	1	1	1	0	0	0	1	1	0	0	0	0	1	0	0
U5	0	0	0	1	1	0	1	0	0	1	0	0	0	1	1
U6	1	1	0	0	1	0	1	1	1	1	0	1	0	0	0
U5	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1

Table 3. The original tag set of mobile video

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
慢歌 (slow song)	抒情 (lyric)	台語 (Taiwanese)	單人 (single)	男 (male)	美聲型 (canto)	華語 (Chinese)	快歌 (fast song)	流行樂 (pop)	粗礦 (crude)	美麗 (beautiful)	火辣 (hot)	可愛 (cute)	搞笑 (funny)	新秀 (rookie)

Table 4. The original tag support of mobile video

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
	慢歌 (slow song)	抒情 (lyric)	台語 (Taiwanese)	單人 (single)	男 (male)	美聲型 (canto)	華語 (Chinese)	快歌 (fast song)	流行樂 (pop)	粗礦 (crude)	美麗 (beautiful)	火辣 (hot)	可愛 (cute)	搞笑 (funny)	新秀 (rookie)
support	0.7	0.7	0.5	0.5	0.5	0.1	1.0	0.3	0.3	0.5	0.3	0.3	0.2	0.2	0.3

Table 5. The screened tag set of mobile video purchasing

	T1	T2	T3	T4	T5	T8	T9	T10	T11	T12	T13	T14	T15
	慢歌 (slow song)	抒情 (lyric)	台語 (Taiwanese)	單人 (single)	男 (male)	快歌 (fast song)	流行樂 (pop)	粗礦 (crude)	美麗 (beautiful)	火辣 (hot)	可愛 (cute)	搞笑 (funny)	新秀 (rookie)
support	0.7	0.7	0.5	0.5	0.5	0.3	0.3	0.5	0.3	0.3	0.2	0.2	0.3

Table 6. The user tag access matrix after tag screening process of mobile video purchasing

	慢歌 (slow song)	抒情 (lyric)	台語 (Taiwanese)	單人 (single)	男 (male)	快歌 (fast song)	流行樂 (pop)	粗礦 (crude)	美麗 (beautiful)	火辣 (hot)	可愛 (cute)	搞笑 (funny)	新秀 (rookie)
U1	1	1	1	1	0	0	0	0	1	0	0	0	0
U2	1	1	1	0	1	1	0	1	0	1	0	0	0
U3	1	1	0	0	1	0	1	1	1	1	0	0	0
U2	0	0	0	1	0	0	1	0	0	0	0	1	1
U1	1	1	0	1	1	0	0	1	1	0	1	0	0
U4	1	1	1	1	0	0	0	0	0	0	0	0	0
U4	1	1	1	0	0	1	0	0	0	0	1	0	0
U5	0	0	0	1	1	0	0	1	0	0	0	1	1
U6	1	1	0	0	1	1	1	1	0	1	0	0	0
U5	0	0	0	0	0	0	0	0	0	0	0	0	1

4.2.2 Tag shrinking phase

Many clustering algorithms are based on independently and identically distributed attributes. Therefore, high dependent tags will affect the performance of clustering algorithm and are often considered as redundant tags. The dependence between two itemsets can be estimated by the confidence of one itemset over the other. From the view of CF-based recommendation systems, the characteristics of users' preference are usually assumed as 1-itemset. Users' preferences should be characterized more precisely instead of 1-itemset, while high dependent tags should be shrunk into attribute for reducing over-concerning to achieve better clustering results. According to the domain considerations and theoretical reasons, we shrink high dependent tags into k-itemsets. The idea is to iteratively find two attributes with high confidence and shrink them into one itemset. Example 2 explains the details of tag shrinking phase.

Example2：Tag shrinking process

Following Example 1, we iteratively compute the confidence of every two attributes and shrink them with high confidence into one itemset. Here, we set the confidence threshold as 0.9, and then the screened tag set is further shrunk by tag shrinking process. Table 7 shows shrunk tag set and the corresponding user purchasing matrix is presented in Table 8. Furthermore, this tag shrinking process will transform the Boolean valued tags into the integer valued itemsets avoiding the potential information loss causing by shrinking.

Table 7. The shrunk tag set of mobile video

(slow song) (lyric) (Taiwanese)	(male) (crude)	(single)	(fast song)	(pop)	(beautiful)	(hot)	(cute)	(funny)	(rookie)

Table 8. The shrunk user tag access matrix of mobile video purchasing

	(slow song) (lyric) (Taiwanese)	(male) (crude)	(single)	(fast song)	(pop)	(beautiful)	(hot)	(cute)	(funny)	(rookie)
U1	3	0	1	0	0	1	0	0	0	0
U2	3	2	0	1	1	0	1	0	0	0
U3	2	2	0	1	1	1	1	0	0	0
U2	0	0	1	0	1	0	0	0	1	1
U1	2	2	1	0	0	1	0	1	0	0
U4	3	0	1	0	1	0	0	0	0	0
U4	3	0	0	1	0	0	0	1	0	0
U5	0	2	1	0	0	0	0	0	1	1
U6	2	2	0	1	1	0	1	0	0	0
U5	0	0	0	1	0	0	0	0	0	1

4.2.3 Tag attribute transformation phase

Although tag screening phase and tag shrinking phase attempt erasing the dependence among tags while remaining sufficient informative tags, it is still difficult to form the i.i.d. and explainable attributes directly from folksonomyt-based tags. Besides, the original purpose of folksonomy-based tags is to describe the properties of content, but not human emotion. Misuse of folksonomy-based tags may be risk of performing poor recommendation results. Take advantage of expertise from psychology which can be represented by tag-preference mapping matrix, the transformation from folksonomy-based tags into taxonomy-based and i.i.d. attributes can be easily built. For maximally utilizing the discriminating information of folksonomy-based tags provide, the dominated itemsets is used to be the middleware of transformation. Example 3 explains the procedure of tag attribute generation phase.

Example3：Tag attribute transformation process

Following the Example 1 and Example 2, user purchasing matrix records the purchase information of 5 users and 15 tags. The illustrated tag-preference matrix is given in Table 9. The weighted transformation is adapted to remain the discriminating information according to the support of each itemset. Table 10 shows the user preference matrix after tag attribute transformation process. The transformation from integer-valued tags to real-valued attributes can dramatically reduce the dimensions of attributes without decreasing the discrimination, since the real-valued tags can remain the same or better expression power.

Table 9. The tag attribute mapping table

	Violent	Pornographic	Interesting	Strength	Evil
Slow song	-1.0	0	1.0	0.5	-0.5
Lyric	-1.0	0.5	0.5	0.5	-0.5
Taiwanese	-0.5	0	0.5	0.5	-0.5
Single	0.5	0	-0.5	-1	-1
male	0	0.5	0.5	1.0	0.5
Fast song	1.0	0.5	0.5	1.0	1.0
pop	-0.5	-0.5	-0.5	-0.5	-0.5
crude	0.5	1.0	0.5	0.5	0.5
beautiful	-1.0	0.5	-0.5	0.5	0.5
hot	-0.5	1.0	-0.5	0.5	1.0
cute	-1	0	-1	0.5	0.5
funny	-1	-0.5	1.0	1.0	-1
rookie	0	-1	0	0.5	0

Table 10. The user preference matrix

	Violent	Pornographic	Interesting	Strength	Evil
U1	-5.65	3.85	1.35	3.5	-2
U2	-2.99	1.01	1	4	-1.5
U3	-0.66	3.34	2.34	4	2
U4	-4.98	1.04	-2.02	3	-3
U5	1	-0.5	2	3.5	0
U6	-1.16	2.84	1.84	4.5	1.5

4.3 Tag attribute generation algorithm

In this section, we will present our tag attribute generation algorithm following the above phases to overcome the curse of dimensionality, form independently and identically distributed attributes and remain the explainable attributes.

Algorithm 1 (attribute generation algorithm)

Input:

(1) User purchasing matrix, $P_{|U|\times|V|}$

(2) Video tagging matrix, $A_{|V|\times|T|}$

(3) Tag-preference mapping matrix, $X_{|T|\times|N|}$

(4) Maximum support threshold, T_M

(5) Minimum support threshold, T_m

(6) Confidence threshold, T_c

Output:

(1) User preference matrix, $M'_{|U|\times|N|}$

Method：

Step 1: Calculate the user tag access matrix $M_{|U|\times|T|}$ by multiplying $P_{|U|\times|V|}$ and $A_{|V|\times|T|}$.

Step 2: For each i, normalize M(:,i).

Step 3: For each tag t_i, screen t_i from T if $S(t_i) > T_M$ or $S(t_i) < T_m$.

Step 4: For each A,B in T, remove A and B from T and add $A \cup B$ into T if $C(A|B) > T_C$.

Step 5: Repeat Step 4 until the cardinality of T remains the same.

Step 6: For each i, $M'(i,j) = \frac{1}{\sum_k M(i,k)} \sum_k M_{i,k} \times f(A_k)$, where $f(A_k) = \frac{1}{|A_k|} \sum_{v \in A_k} X(v,j)$.

Chapter 5 Experiment

5.1 Experimental Design

In this thesis, we aim to generate an i.i.d. attribute set which is proper to characterize the users' preferences while conquering the curse of dimensionality. To evaluate our attribute generation algorithm, the entertainment video recommendation system in e-commerce is selected as application. The experiment data were offered by a commercial entertainment video Web shop in Taiwan. There are total 1487 available videos and purchase information of 10906 customers during 2008/06/12 to 2009/05/07. Each video contains several suitable tags from 117 predefined tags. The 28249s transactions are separated into training data (from 2008/06/12 to 2008/11/23) and testing data (from 2008/11/24 to 2009/05/07), The training data is further split into training set and validation set for the purpose of model selection. Figure 3 shows the data sizes and the corresponding duration and the detail of data statistics is shown in Table 11.

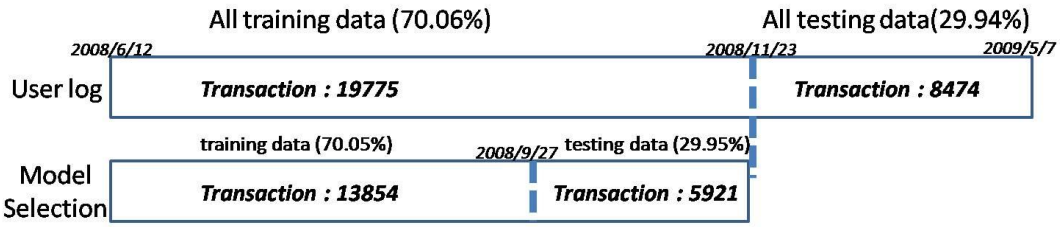


Figure 3. The data set of all user log and Model Selection

Table 11. The detail of data statistics of purchase information

	Transaction Number (Percentage)		Customer Number (Percentage)		Average Purchase Per Customer Mean(Std.)	
Training Data	19775 (70.06%)	13854 (49.04%)	8509 (78.02%)	5869 (53.81%)	2.32(3.71)	2.36 (1.87)
		5921 (20.96%)		2640 (24.21%)		2.43 (4.08)
Testing Data	8474(29.94%)		3454(31.67%)			

Our attribute generation algorithm consists four phases: tag screening phase, tag shirking phase, tag enhancing phase and tag transformation phase. In tag screening phase, only tags with support between 0.1 and 0.9 are kept. While deleting tags with confidence over 0.9 in tag shirking phase, only tags with support over 0.5 are kept in tag enhancing phase. In tag transformation phase, the transformation is based on the tag attribute mapping table in Table 9. Each phase will form different characterizations of users' preferences.

For evaluating the effects of different characterizations of users' preferences, we use content-based collaborative filtering (CBCF) recommendation approach based on the assumption that people will interest in those persons with their similar preferences interest in, together with well-known K-means clustering algorithm [23] to cluster users with similar preferences. K-means starts with random k data points as centroids; clusters data points with k centroids and regenerate k centroids from the corresponding clusters. The algorithm terminates while centroids are the same as before. The similarity measures, Euclidean distance and cosine similarity, are selected for K-Means according to different types of attributes, continue and discrete, respectively. In particularly, for two vectors V_1 and V_2, the Euclidean distance of V_1 and V_2, $d(V_1, V_2) = \sqrt{(V_1 - V_2)^2}$, and the cosine similarity of V_1 and V_2, $Cos(V_1, V_2) = \frac{V_1 \cdot V_2}{|V_1||V_2|}$. The length of recommendation lists are set with two version 5

recommendations and 10 recommendations. The goals of experiments are to evaluate (1) the effects of dimension reduction, (2) the effects of generated attributes on charactering users' preference, and (3) the effects of generated attributes on recommendation. The experiment results will be analyzed and discussed below.

5.2 The effects of dimension reduction

The experiments are run on a desktop computer with 2.3GHz Intel Core Quad CPU , 4Gb RAM and Microsoft Windows XP SP3 operating system. The algorithm is implementation by Microsoft Visual Studio 2005 and Microsift SQL server 2005. Each phase of proposed algorithm will reduce the dimension of experimental data. The tag screening phase will reduce 117 tags (1-itemset without screening) to 44 tags (1-itemset with screening). While reducing 44 tags to 17 tags (closed itemset) in the tag shirking phase, 17 tags will further be reduced into 15 tags (large closed itemset) in the tag enhancing phase. After the tag transformation phase, 5 human emotional attributes will be generated. The obvious advantages of smaller attribute number are faster processing time and smaller required memory storage. Table 12 shows the processing time and required memory storage with different attribute set.

Table 12. The performance of different attribute sets

Attribute set	Process Time (毫秒)		Memory space	Attribute count
	推5片	推10片		
1-itemset (without creening)	343.906	422.969	1130.58KB	114
1-itemset (with screening)	240.750	360.969	425.18KB	44
Closed itemset	217.656	220.469	164.27KB	17
Large closed itemset	211.875	186.094	144.95KB	15

The original one itemset are built from folksonomy-based, there are many redundant tags which are low-informative such as the tag with excessively high support such as “Chinese(華語)” or excessively low support such as “Canto(美聲型)”. The low-informative tags were helpless to distinguish different characteristics. Reducing low-informative tags and shirking high dependence tags can improve processing time and decrease required memory storage. Traditional dimension reduction techniques may suffer the trade-off between attribute number and performance. However, we argue that our proposed attribute generation algorithm can generate smaller attribute set, while improving the performance of recommendation system and hence conquering the curse of dimensionality.

5.3 The effects of generated attributes on charactering users’ preferences

Different attribute sets describe distinct user’s preference characterization by the same characterization algorithm, K-means algorithm used here. One issue of K-means algorithm is how to decide the number of clusters. The common solution is via model selection. Since the purpose of this thesis is to build a recommendation system, the criterion evaluating the performance is selected as hit rate. We call hit if one of someone’s testing data is in the recommendation list. So the hit rate means the fraction of hit number divided by the number of total testing users. The recommendation list is built personally by half old video contents and half new video contents which do not have purchase history. The priority of old video lists is in decreasing order of the purchase number of videos purchased by the users in the same cluster. New video content list is built by the similarity between new video and cluster characteristics.

For each attribute set and different recommendation list length, we validate each

model with cluster number from 2 to 10. The model with the best hit rate in the validate set is selected. The more cluster number represents the more precisely characterization of users' preferences. Figures 4, 5, 6 and 7 show the results of model selection on different attribute sets, 1-itemset without screening, 1-itemset with screening, closed itemset and large closed itemset, respectively.

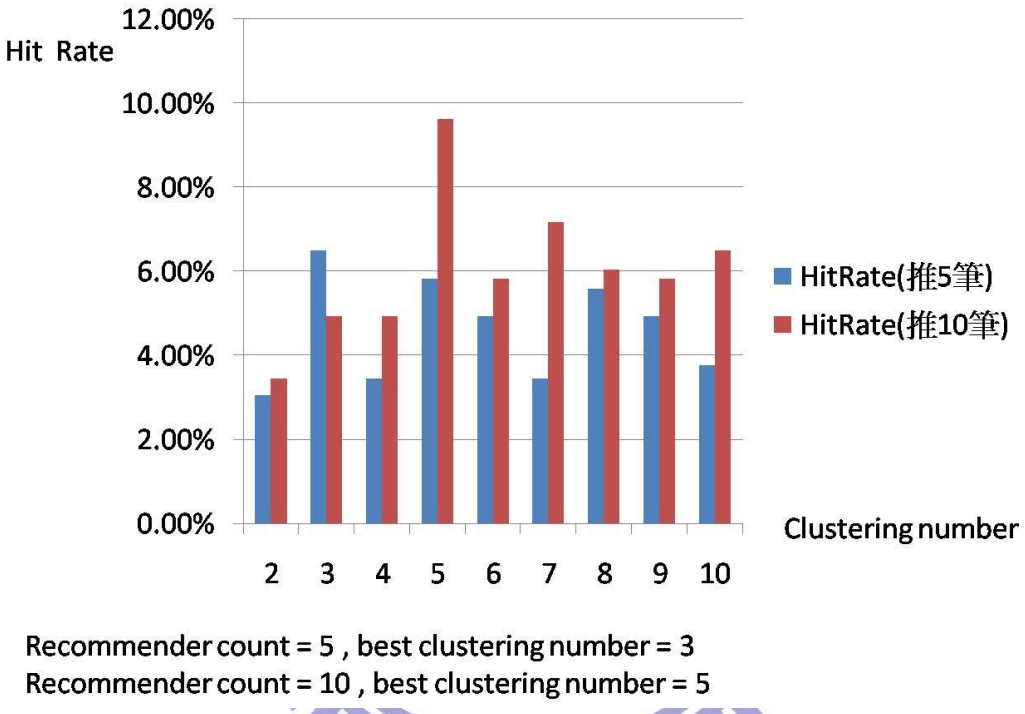
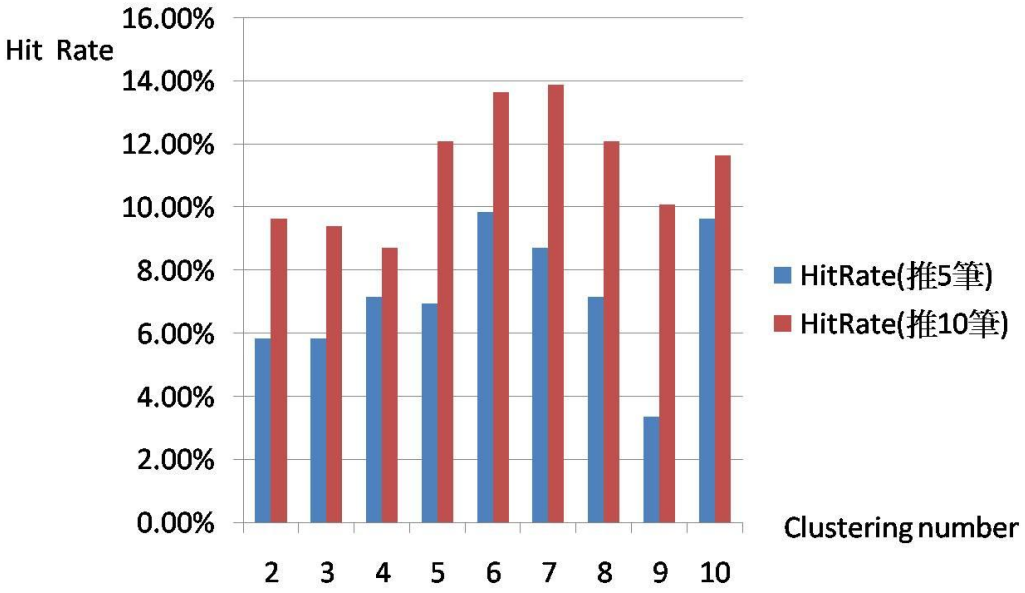
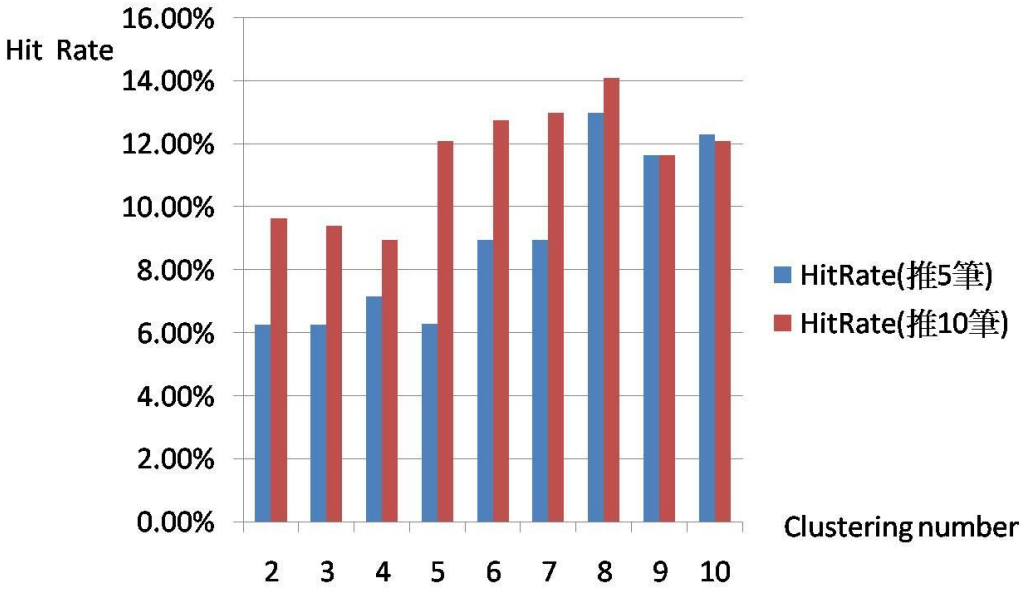


Figure 4. 1-itemset without screening



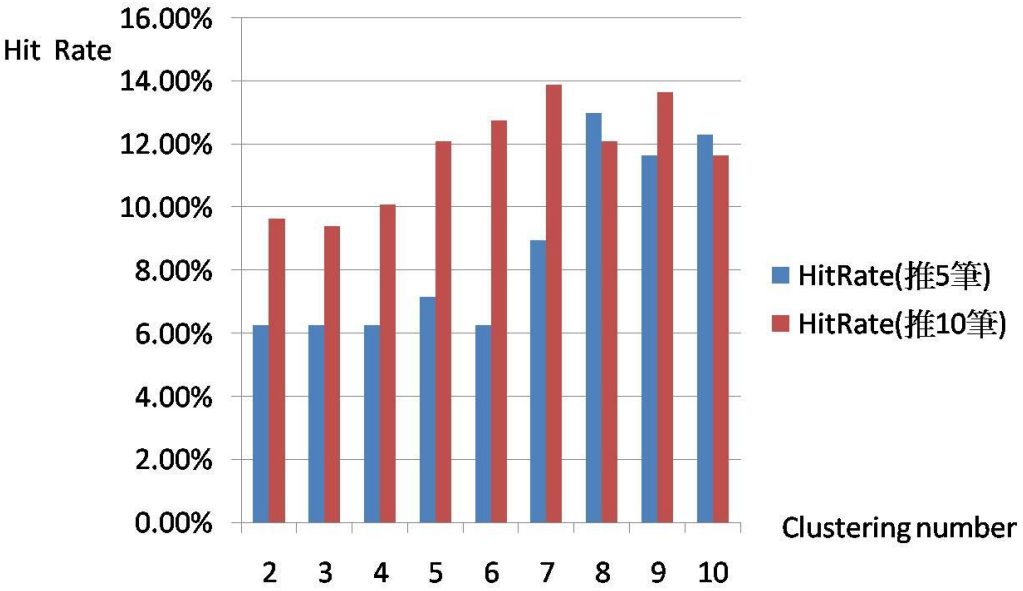
Recommender count = 5 , best clustering number = 6
Recommender count = 10 , best clustering number = 7

Figure 5. 1-itemset with screening



Recommender count = 5 , best clustering number = 8
Recommender count = 10 , best clustering number = 8

Figure 6. Closed itemset with screening



Recommender count = 5 , best clustering number = 8
Recommender count = 10 , best clustering number = 7

Figure 7. Large closed itemset with screening

In this experiment, the best cluster numbers of 1-itemset without screening, 1-itemset with screening and closed itemset are in increasing order whenever the recommendation list length is 5 or 10. This shows that the tag screening phase and the tag shirking phase can form attributes which are better for charactering users' preferences. The tag enhancing phase does not perform well here. The philosophy of the tag enhancing phase is that user purchases this video because he interests in. But the user does not purchase this video cannot provide reciprocally information. Therefore, ignoring lower informative attributes (support $<$ 0.5) will slightly decreases the describing power of attributes. However, this effect is not obviouse in our experiments.

5.4 The effects of generated attributes on recommendations

The goal of recommendation system is to recommend products which customers will interest in and purchase. The hit rate is a suitable criterion to measure the performance of a recommendation system. Our recommendation is based on Content-Based Collaborative Filtering recommendation approach with K-means clustering algorithm. However, K-means algorithm is very sensitive on initial setting. The first randomly selected k centroids will affect the performance K-means algorithm much. To avoid this issue of K-means, we repeat our experiments 50 times on each attribute set. For each experiment, the number of clusters is decided by the above model selection results. The experimental result is shown in Table 13. The results show that our generated attributes perform well in recommending entertainment videos; especially the tag screening phase improves the most hit rate. This explains that low informative attributes have bad influences on recommending video. The tag enhancing phase improves performance when the recommendation list

length is 5. This is because small recommendations need to focus on the sufficient users' preferences.

Table 13. The results of recommendations

Attribute set	Recommender count:5		Recommender count:10		method	Attribute count
	Best k in model selection	AVG hit rate(std)	Best k in model selection	AVG hit rate(std)		
1-itemset (without screening)	3	4.080% (0.013)	5	10.885% (0.029)	Original tag set	117
1-itemset (with screening)	6	6.114% (0.014)	7	11.570% (0.015)	0.1 < support < 0.9	44
Closed itemset (with screening)	8	6.930% (0.008)	8	11.954% (0.025)	confidence < 0.9	17
Large Closed itemset (with screening)	8	7.262% (0.008)	7	11.598% (0.018)	0.5 < Sup < 0.9	15

In our experiment, we are interested in the characteristics between different clusters, for example, the characteristic "Idol(偶像型) and 專輯(Album)" and "Idol(偶像型) and Female(女)" are different clusters' characteristic. According to our observations, "Idol(偶像型)" is the characteristic of single cluster, we can use the more precisely attributes to describe the users' preferences and hence clustering customers with similar preferences more precisely. Our experimental results demonstrate that out proposed attribute generation algorithm perform well on the entertainment video recommendation in three aspects: (1) faster processing time and smaller memory storage, (2) better characterization of users' preferences, and (3) better performance of recommendation system.

Chapter 6 Conclusion and Remarks

Many clustering approaches assume that the clustering attributes are well-defined; however, nowadays, folksonomy-based tagging is more popular for users. So there are many noises in the original attribute or tags would influence the clustering result. In this situation, how to extract significant attributes from folksonomy to taxonomy is of most importance.

In our attribute extraction method, in the tag screening algorithm, we can screen the noise at the beginning. Second, to solve the tag dominated problem, we propose the tag shrinking algorithm to shrink the dominated tags and complement their weights; therefore we can find not only explicit characteristics but also implicit characteristics and improve the social recommendation systems. After that, there are still some un-shrunk tags may raise sparsity problem. In order to solve this problem, we propose an attribute transformation process to generalize the sparse tag value with similar semantic meaning.

According to the analysis of the implicit characteristic on different cluster, we can design adaptive recommendation strategy to achieve better results.

Reference

[1] Adomavicius, G. and A. Tuzhilin, "*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, 2005. **17**(6): p. 734 - 749.

[2] B.N. Miller, I. Albert, S.K. Lam, J.A. Konstan, and J. Riedl, "*MovieLens Unplugged: Experiences with an Occasionally Connected Recommender System*," Proc. Int'l Conf. Intelligent User Interfaces, 2003.

[3] Chia-Chuan Hung, Yi-Ching Huang and Jane Yung-jen Hsu, et al., "*Tag-Based User Profiling for Social Media Recommendation"*, in Workshop on Intelligent Techniques for Web Personalization and Recommender Systems. 2008

[4] Charu C. Aggarwal, Jiawei Han, Jianyong Wang and Philip S. Yu, "*A Framework for Projected Clustering of High Dimensional Data Streams*"

[5] Charu C. Aggarwal, Cecilia Proconpiuc, Joel L. Wolf, Philip S. Yu and Jong So Park, "*Fast Algorithms for Projected Clustering*", Proceeding of ACM SIGMOD Conference Philadelphia PA, pp. 61-72, 1999.

[6] D. Billsus, C.A Brunk, C. Evans, B. Gladish, and M. Pazzani, "*Adaptive Interfaces for Ubiquitous Web Access*," Comm. ACM, Vol. 45, No.5, pp. 34-38, 2002

[7] G. Linden, B. Smith, and J. York. "*Amazon.com Recommendations: Item-to-Item Collaborative Filtering*," IEEE Internet Computing, Jan/Feb. 2003.

[8] Hauptmann, A., et al., "*Can High-Level Concepts Fill the Semantic Gap in Video Retrieval? A Case Study With Broadcast News"*. IEEE Transactions on Multimedia, 2007. **9**(5): p. 958-966.

[9] J. Ben Schafer, Joseph Konstan, John Riedi."*Recommender systems in e-commnerce*", ACM Special Interest Group on Electronic Commerce. 1999

[10] J.Bobadilla, F.Serradilla, A. Hernando, MovieLans, "*Collaborative filtering adapted to recommender systems of e-learning*", Knowledge-Based System. 2009

[11] Kasun Wickramaratna, Student Member, IEEE, Miroslav Kubat, Senior Member, IEEE, and Kamal Premaratne, Senior Member, IEEE, "*Predicting Missing Items in Shopping Carts*" IEEE TRANSCACTION ON KNOWLEDGE AND DATA ENGINEERING, Vol. 21, No. 7, July 2009.

[12] Lance Parsons, Ehtesham Haque and Huan Liu, "*Subspace clustering for high dimsnsional data: a review*", ACM SIGKDD Explorations Newsletter Vol. 6, 2004

[13] Lun-ping Hung, "*A personalized recommendation system based on product taxonomy fro one-to-one marketing online*" Expert Systems with Application 2005 p. 383-392

[14] M. Pazzani, "*A Framework for Collaborative, Content-Based, and Demographic Filtering"*, Artificial Intelligence Rev. pp. 393-408, Dec. 1999"

[15] M.Balabanovic and Y. Shoham, *"Fab: Content-Based, Collaborative Recommendation,"* Comm. ACM, Vol. 40, No. 3, pp. 66-72. 1997

[16] Milan Vojnovic, James Cruise, Dinan Gnawardena, and Peter Marbach, "*Ranking and Suggesting Popular Items*", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 21, No. 8, AUGUST 2009.

[17] Milind Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann and Jon Curtis, *"Large-Scale Concept Ontology for Multimedia"*, in IEEE Multimedia. 2006. p. 86-91.

[18] N.Belkin and B. Croft, "*Information Filtering and Information Retrieval* **:** *two sides of the same coin?*" Comm. ACM, Vol. 35, No. 12, pp. 29-38, 1992

[19] P. Melville, RJ. Moony, and R. Nagarajan, "*Content-Boosted Collaborative Filtering for Improved Recommendations*," Proc. 18th Nat'l Conf. Artificial Intelligence. 2002.

[20] R.Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrival*. Addison-Wesley, 1999

[21] Robert Wetzker, Winfried Umbrath and Alan Said, ”*A hybrid approach to item recommendation in folksonomies*”, In Proceedings of ESAIR ’09, pp. 25-29, New York, NY, USA, 2009. ACM.

[22] Snoek, C.G.M., et al., “*Adding Semantics to Detectors for Video Retrieval”*. IEEE Transactions on Multimedia, 2007. **9**(5): p. 975-986.

[23] Tapas Kanungo, Senior Memberm, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, Senior Member, IEEE, “*An Efficient k-Means Clustering Algorithm: Analysis and Implementation*”, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLGENCE Vol. 24, No. 7, JULY 2002.

[24] Wang, F.-h. and S.-y. Jian, “*An Effective Content-based Recommendation Method for Web Browsing Based on Keyword Context Matching”* Journal of Informatics & Electronics, 2006. **1**(2).

[25] Wujian Yang, Zebing Wang and Mingyu You, ”*An improved Collaborative Filtering Method for Recommendation’s Generation*”, IEEE International Conference on Systems, 2004.

[26] Yi-Fan Wang, Yu-Liang Chuang, Mei-Hua Hsu and Huan-Chao Keh, et al., “*A personalized recommender system for the cosmetic business”*. Expert Systems with Applications, 2004. 26(3): p. 427-434.