パターン分析を利用した動画リコメンド仮説

大阪府立大学　経済学部
経営学科　4年生
片岡弘貴

報告概要

- 本報告では，レコメンデーションコンテスト2009より提供された，ある動画視聴サイトの視聴およびお気に入り登録データを用いて，あるユーザに対するお気に入りに入ると予想される動画を推奨する方法を提案する．
- 分析では，様々な問題から既存手法の適用が困難であるため，動画タイトルをTMSを利用して工夫することでユーザが求める分野の動画を絞り込むとともに，お気に入りの既存の登録履歴のパターンをVMSによって計算し，この2つの観点から新たなレコメンド手法を提案する．
- 与えられているデータから，計算結果によって，提案手法の有効性を確認する

分析目的

- 動画視聴サイト「サグールテレビ」におけるユーザーに対してより好むと思われる動画を推薦すること。

データ提供元：http://www.team-lab.com/

分析課題

- レコメンデーションコンテスト2009*より提供されたデータは2008/1/29～2009/5/8の間に
  各ユーザーが登録したお気に入りの動画の情報と、2009/2/7～2009/5/8におけるユーザーの行動履歴（動画の視聴や検索履歴）となります。

課題!!

「お気に入り」に動画を登録している数が20個以上のユーザーのお気に入りデータからユーザー毎にランダムに動画情報が10個削除されています。

その削除された10個が何かを推測すること！

*レコメンデーションコンテスト2009：http://kgmod.jp/contest/

提供データ紹介(1)

■動画マスターデータ(1,780,463件)

サンプル

mid	title	url	turl	sec	site	delete	uname
100	AAA	http:____	http:---	300	youtube	0	name

- mid: 動画識別ID
- title: 動画タイトル
- url: 動画URL
- turl: サムネイルURL
- sec: 動画の時間
- site: 動画掲載サイト
- delete: 削除されているか
- uname: アップロード者の名前

■お気に入り登録データ

サンプル

uid	mid	cdate
1	100	20090101

- uid: ユーザー識別ID
- cdate: 登録日

お気に入り登録数20個以上のユーザーのデータ(32,083件) → 448人

お気に入り登録数20個未満のユーザーのデータ(7,881件) → 2,016人

注!!これ以降出てくる「お気に入り登録データ」という言葉は
基本的に推奨すべき登録数20個以上のユーザーデータを指します

提供データ紹介(2)

■動画再生履歴データ

サンプル

ユーザーのセッションID

再生日＋時刻

uid	sid	ip	mid	date
1	*****	*****	100	20090101120000

ユーザーIPアドレスのMD5ハッシュ

■動画中断履歴データ

中断日＋時刻

経過時間(秒)

uid	sid	ip	mid	date	sec
1	*****	*****	100	20090101120030	30

■動画完了履歴データ

再生完了日＋時刻

uid	sid	ip	mid	date
1	*****	*****	100	20090101120500

■動画検索履歴データ

検索ワード

uid	keyword	cnt
1	A	10

検索回数

サンプル

お気に入り登録数に関係なくすべてのユーザー分のファイル

先のお気に入り登録データ同様に、登録数20個以上のユーザーのファイルと20個未満のファイルとに分かれている。
また、各ユーザーごとにお気に入りから削除された動画に関する行動履歴は削除されている

課題の難しさ

お気に入り登録動画数20個以上のユーザー448人に対して、抜かれたであろう動画を各ユーザーに対して10個ずつ列挙する。

注：抜かれた動画は必ず動画マスターデータに存在するものである

単純に動画マスターデータからランダムに列挙すると
10/1,780,463＝約0.00056%

この確率をいかにして上げるかが問題である

分析する際の問題点

1. 行動履歴データ期間の問題

- お気に入り登録データと行動履歴データの期間の違い
- 上記に起因する可能性のある行動履歴データが存在しないユーザーの問題

2. サイト独自のリコメンドシステムの問題

- サイト特有の方法に起因する問題

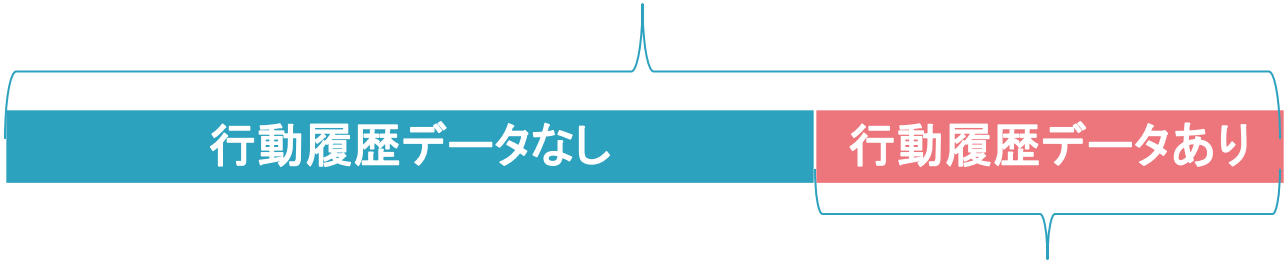
3. 評価値が存在しない問題

- サイト固有の問題ではないが、すべてのユーザーに存在するお気に入り登録データは、登録有無を示す1,0のデータのため、評価値を想定したような協調フィルタリングには不向き

行動履歴データ期間の問題

データの記録期間が異なるため，行動履歴データを十分に使えない

お気入り登録期間は2008/1/28～2009/5/8



行動履歴は2009/2/7～2009/5/8

実際に動画を推奨するべき448人の中で行動履歴のデータを持つ人は約200人*である。

*行動履歴データの種類によってデータ人数が変わるため

まったく行動履歴のないユーザーも存在するのでそれを入力とする決定方法は難しい

サイト独自のリコメンドシステムの問題

- 「サグールテレビ」のサイトではブラウザ上で動画再生部分と動画検索部分が独立しており、動画を視聴しつつ動画を検索できる。
- ユーザーは視聴したい動画を再生画面にドラッグ＆ドロップすることで視聴予定の動画を溜めることができる。(現在視聴動画が終了次第次が再生される)
- 視聴予定の動画がなかった場合、自動的にサイトがリコメンドする動画が再生される。

このサイト独自のシステムによって再生された動画がユーザーが自身で再生予定に入れた動画なのか、サイトが自動的にリコメンドした動画なのか判断ができない!!

データの問題を踏まえたアプローチ

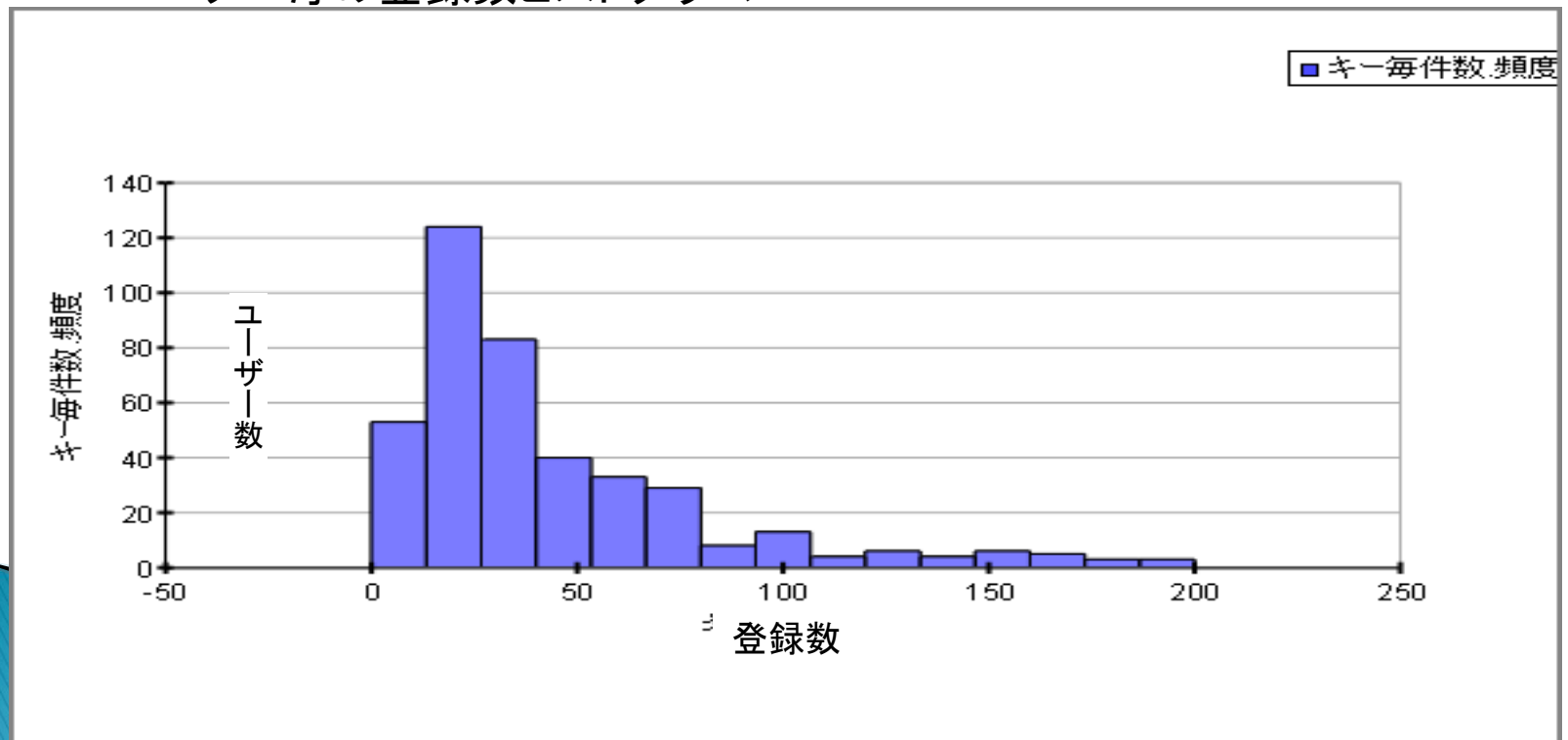
- 協調フィルタリングを使うことなく、別の方法を考案する。この際データのないユーザーが存在する行動履歴データよりも全員のデータが存在するお気に入り登録データをメインで使用する。

動画を推奨すべきユーザーの
お気に入り登録データの分析を行う

お気に入り登録データ基礎分析

- ユーザー数　・・・448人
- データ件数　・・・32,082件
- ユーザーあたり平均動画登録数　・・・71.6件

ユーザー毎の登録数ヒストグラム



お気に入り登録データのテキストマイニング

- 単語頻度解析

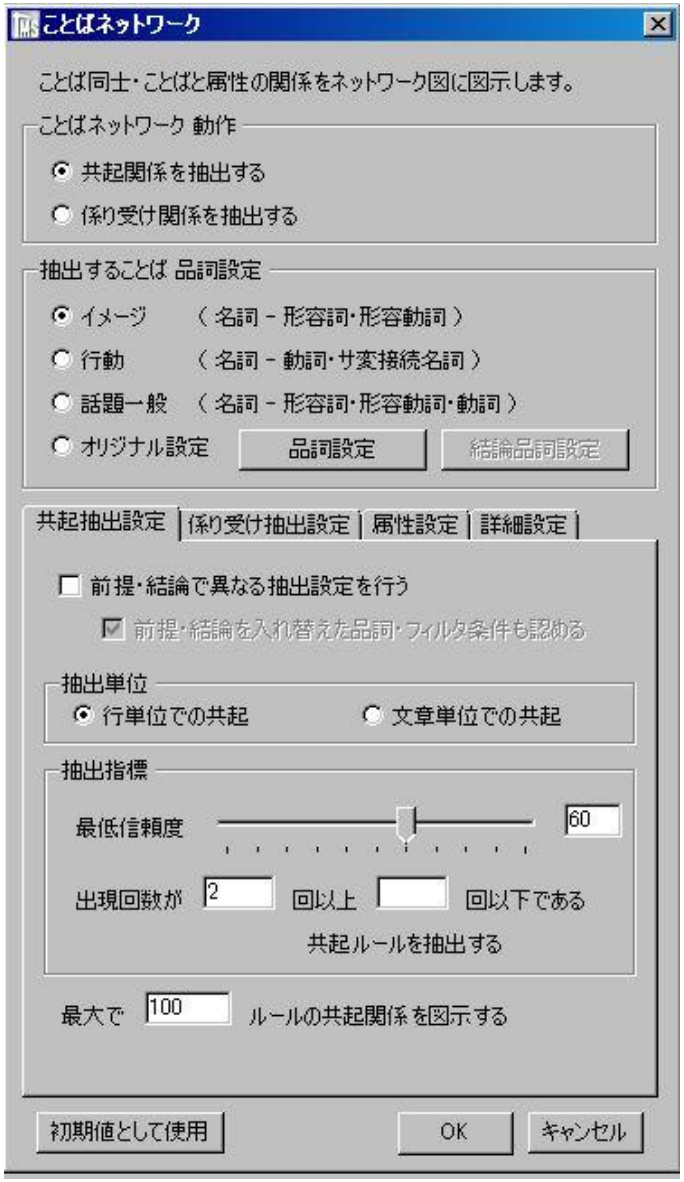
お気に入りに登録されている動画のタイトルで単語頻度解析を行ってみると比較的音楽に関するワードの件数が多いことがわかる

単語	品詞	頻度
ちる	動詞	482
ＰＶ	名詞	456
1	名詞	330
2	名詞	314
Ｌｉｖｅ	名詞	235
いる	動詞	138
ｌｉｖｅ	名詞	134
初音ミク	名詞	115
ＬＩＶＥ	名詞	111
母	名詞	110
腐る	動詞	106
放る	動詞	106
坂本真綾	名詞	100
Ｐｅｒｆｕｍｅ	名詞	97
椎名林檎	名詞	94
木村	名詞	92
恋	名詞	80
花	名詞	78
愛	名詞	76
尾崎豊	名詞	76

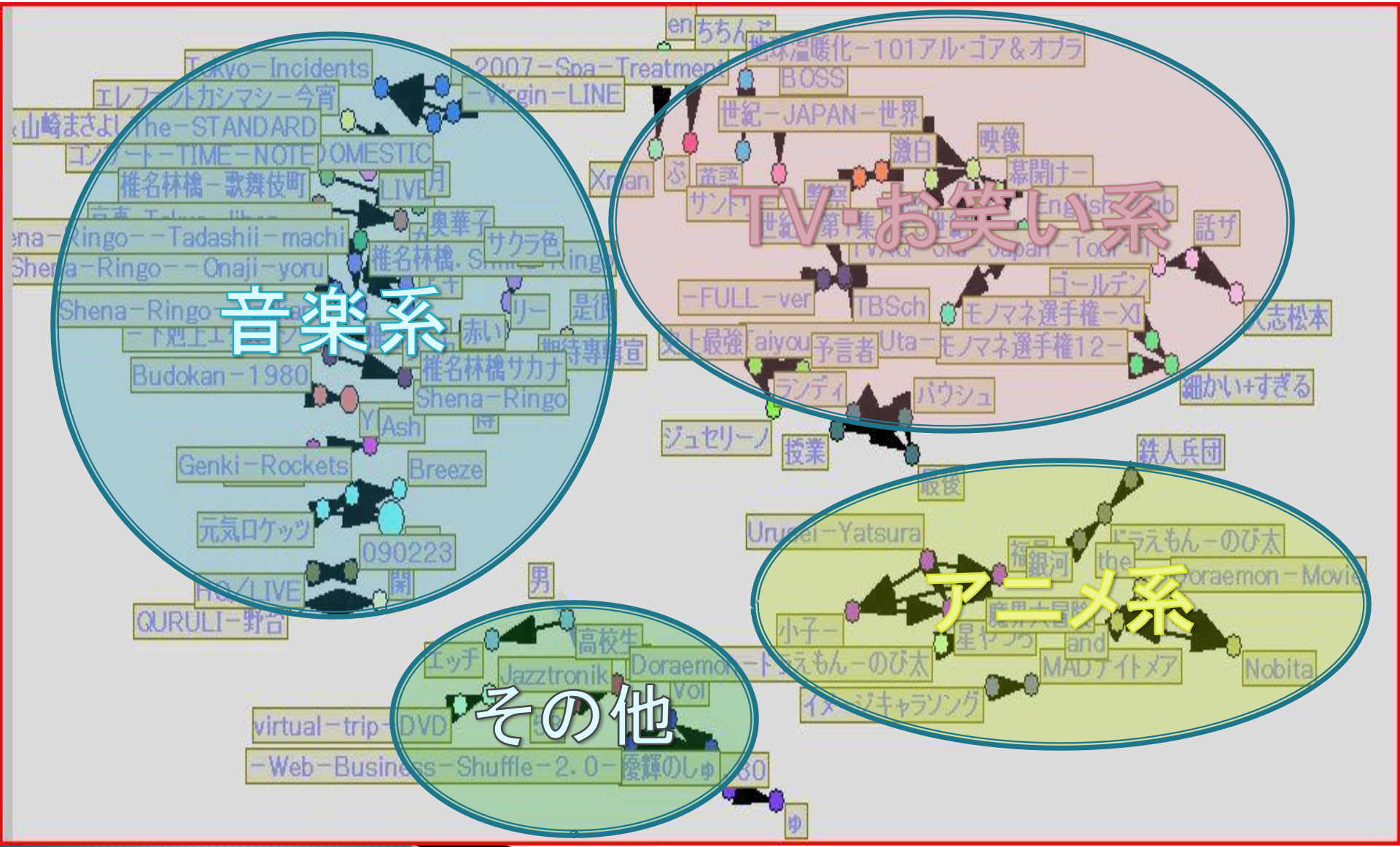
TMSの活用

- TextMiningStudioにある話題分析「ことばネットワーク」を用いて、共起関係にある言葉の抽出をする。

右図の数値設定で解析し、ジャンルごとに言葉を分類する。



お気に入りタイトルによることばネットワーク解析結果



「音楽系」動画の決定方法

カラオケ「JOYSOUND」*に登録されている約2万種類のアーティスト名を動画マスターデータのタイトルから検索し、その検索に合致したものを音楽系動画と定義する。

466,979件の動画を検索

これを「音楽系」動画とする

同様に、TV・お笑い系、アニメ系、その他と分類した

今回は単語頻度解析、ことばネットワークともに最も多く検出された「音楽系」のジャンルに絞って動画を推奨する

*JOYSOUND：http://joysound.com/ex/search/karaoke/index.htm

音楽系動画とお気に入りの関係

		お気に入りに登録	
		YES	NO
音楽系動画であるか	YES	15，099件	451，880件
	NO	16，983件	1,296,501件

仮説） 音楽動画かそれ以外かとお気に入りに入るか入らないかには関連がない
(χ2乗値＝7330.32…)⇒棄却

ユーザーは音楽系動画をお気に入りに登録しやすい傾向がある!!

アーティスト名に関して

- 音楽系動画がお気に入りに登録されやすいことはわかったが、ユーザーごとに好みがあるはずなので、さらにアーティスト名ごとに分析をする。

- JOYSOUNDに登録されているアーティストは20,266件あるが、コラボ企画などで、一つのアーティスト名に複数のアーティスト名が入っているケースがある。

  例：EXILE＆倖田來未　など

お気に入りに関連しユニークなアーティスト名一覧を作成（7,198アーティスト）

ユーザーの嗜好分析

- 「音楽系」動画に関心のあるユーザーを選択する。

  各ユーザーがお気に入りに登録している動画のうち、音楽系動画の占める割合が、動画マスターデータにおける音楽系動画の割合(26.23%)よりも高いユーザーを選択する。

448人→348人

この348人に対して音楽系動画を推奨したい!!

動画の推奨基準

1. ユーザーがお気に入り登録してる音楽系動画におけるアーティストの割合
2. アーティスト名が含まれる動画全体におけるお気に入りに登録された割合
3. ユーザーごとに推奨アーティスト数とその動画の数を変更
4. アーティストベースのパターン分析による類似ユーザーの可能性

基準1.ユーザーごとのアーティストの割合

- ユーザーがお気に入りに登録している音楽系動画に何アーティストを含むかはユーザー次第である。
- 登録している動画に対して割合の高いアーティストの動画を優先して推奨する。

基準2.推奨しやすいアーティストの決定

- アーティスト名で動画マスターから検索したときに検索される動画に対してお気に入りに登録される割合を計算する。
- 割合が高いほど推奨する際の選択肢が減ることになる。

基準3.10個の動画の割り当て

- 基準1で計算した割合からドント方式で10個を割り当てる。ただし、複数の中から一つを選択する際は基準2で計算した割合の高いアーティストを優先して割り当てる。以下のサンプルの場合Aから2個、Bから2個、C～Hまでは1個ずつ動画を選択する。

サンプル	アーティスト名	A	B	C	D	E	F	G	H	I	J	K
	動画件数	10	7	6	6	5	5	5	4	3	3	3
	基準1	0.175439	0.122807	0.105263	0.105263	0.087719	0.087719	0.087719	0.070175	0.052632	0.052632	0.052632
決定順序	1	0.175439	0.122807	0.105263	0.105263	0.087719	0.087719	0.087719	0.070175	0.052632	0.052632	0.052632
	2	0.087719	0.122807	0.105263	0.105263	0.087719	0.087719	0.087719	0.070175	0.052632	0.052632	0.052632
	3	0.087719	0.061404	0.105263	0.105263	0.087719	0.087719	0.087719	0.070175	0.052632	0.052632	0.052632
	4	0.087719	0.061404	0.052632	0.105263	0.087719	0.087719	0.087719	0.070175	0.052632	0.052632	0.052632
	5	0.087719	0.061404	0.052632	0.052632	0.087719	0.087719	0.087719	0.070175	0.052632	0.052632	0.052632
	6	0.057895	0.061404	0.052632	0.052632	0.087719	0.087719	0.087719	0.070175	0.052632	0.052632	0.052632
	7	0.057895	0.061404	0.052632	0.052632	0.04386	0.087719	0.087719	0.070175	0.052632	0.052632	0.052632
	8	0.057895	0.061404	0.052632	0.052632	0.04386	0.04386	0.087719	0.070175	0.052632	0.052632	0.052632
	9	0.057895	0.061404	0.052632	0.052632	0.04386	0.04386	0.04386	0.070175	0.052632	0.052632	0.052632
	10	0.057895	0.061404	0.052632	0.052632	0.04386	0.04386	0.04386	0.035088	0.052632	0.052632	0.052632

注：各アーティストの基準2の値はアルファベットが若いほど高いとする

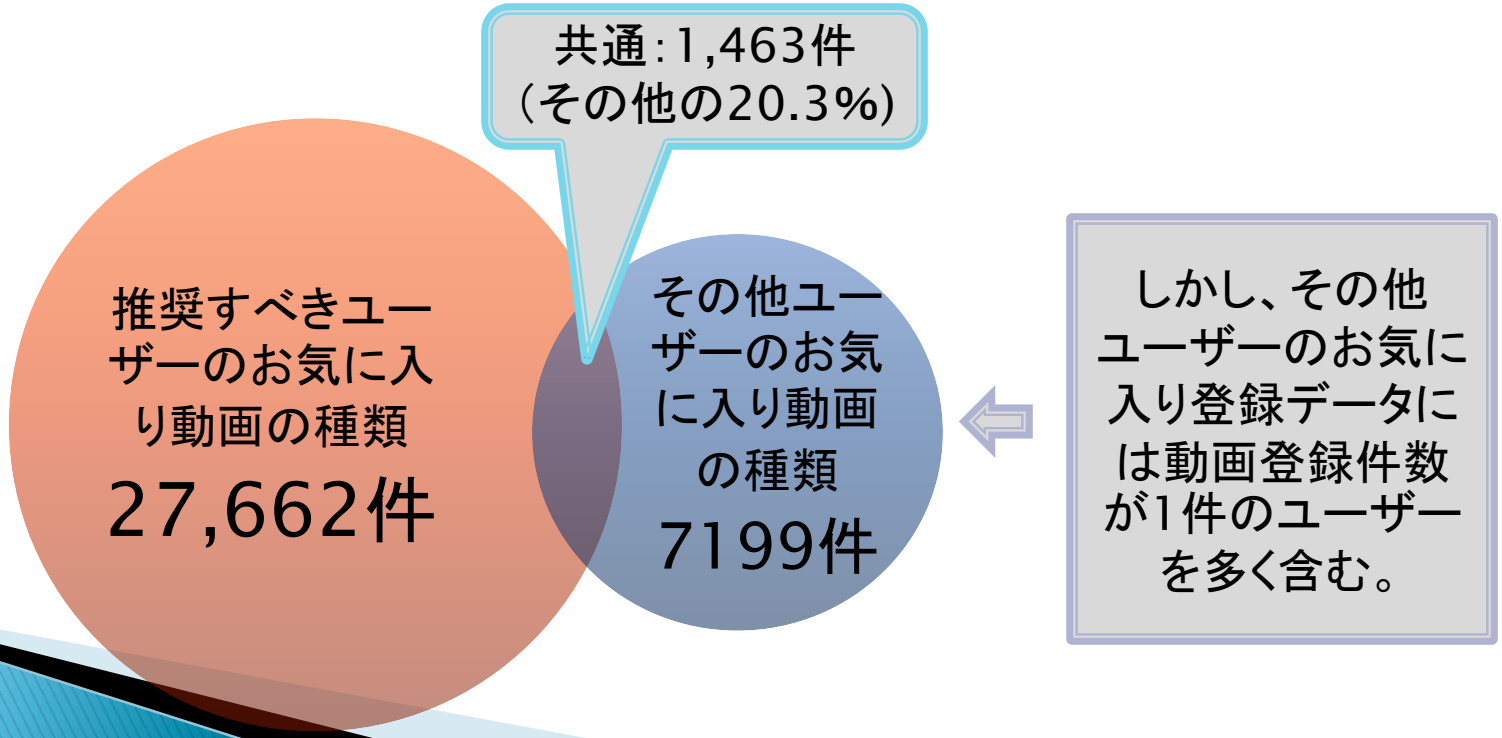
基準4.パターン分析による類似ユーザーの可能性

- 各アーティストの推奨動画の個数を決定した後の動画の決定に際して、アイテムベースのアソシエーション分析を行うことで、類似ユーザー群を見つけ出し、そのユーザー群の中でお気に入りに入れられている同一アーティストの動画を優先的に選択する。

なぜ、他ユーザーの登録している動画を優先するのか？

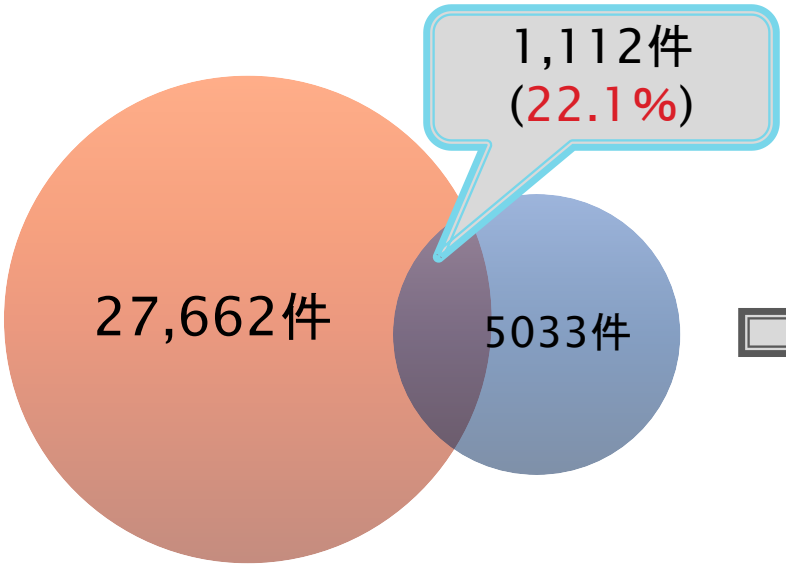
お気に入りに入る動画の種類と共通性

- 今回お気に入り登録データは動画を推奨すべき448人のデータの他に、お気に入り登録20個未満のユーザーのデータがある。この二つの登録データの共通動画の個数を見ると以下のようになる。

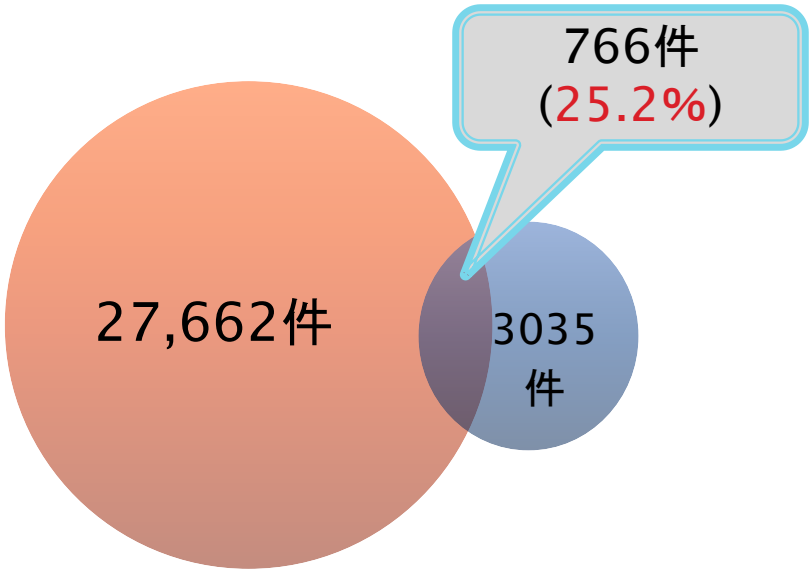


複数登録者の共通性

その他ユーザー
お気に入り登録数5件以上の場合



その他ユーザー
お気に入り登録数10件以上の場合



以上より、お気に入り動画登録件数が多いと
他のユーザーと同じものを登録する確率が上昇する。
(χ2乗値＝10.48)

類似ユーザーの選定

- アーティスト名ベースでのアソシエーション分析を行い、共起されやすいユーザーIDを見つけ、そのパターンを類似ユーザーのグループとする。

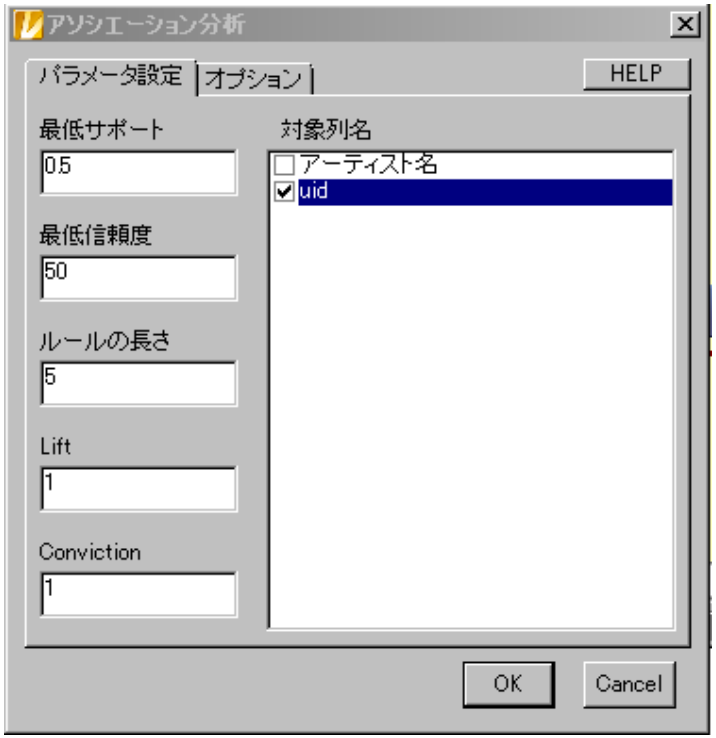
キー列

対象列

データ形式

	アーティスト名	uid
1244	GLAY	2756
1245	GLAY	3303
1246	GLAY	5361
1247	GLAY	9120
1248	GLAY	9366
1249	GO!GO!7188	24
1250	GO!GO!7188	353
1251	GO!GO!7188	421
1252	GO!GO!7188	4185
1253	GO-2	52
1254	GO-2	1155
1255	GOING-UNDER	1973
1256	GOING-UNDER	2214
1257	GOING-UNDER	3845
1258	GOING-UNDER	6332
1259	GReeeeN	32
1260	GReeeeN	35
1261	GReeeeN	54
1262	GReeeeN	301
1263	GReeeeN	421

パラメータ設定を以下の数値にする



アソシエーション分析結果

リフト値で降順にソートする

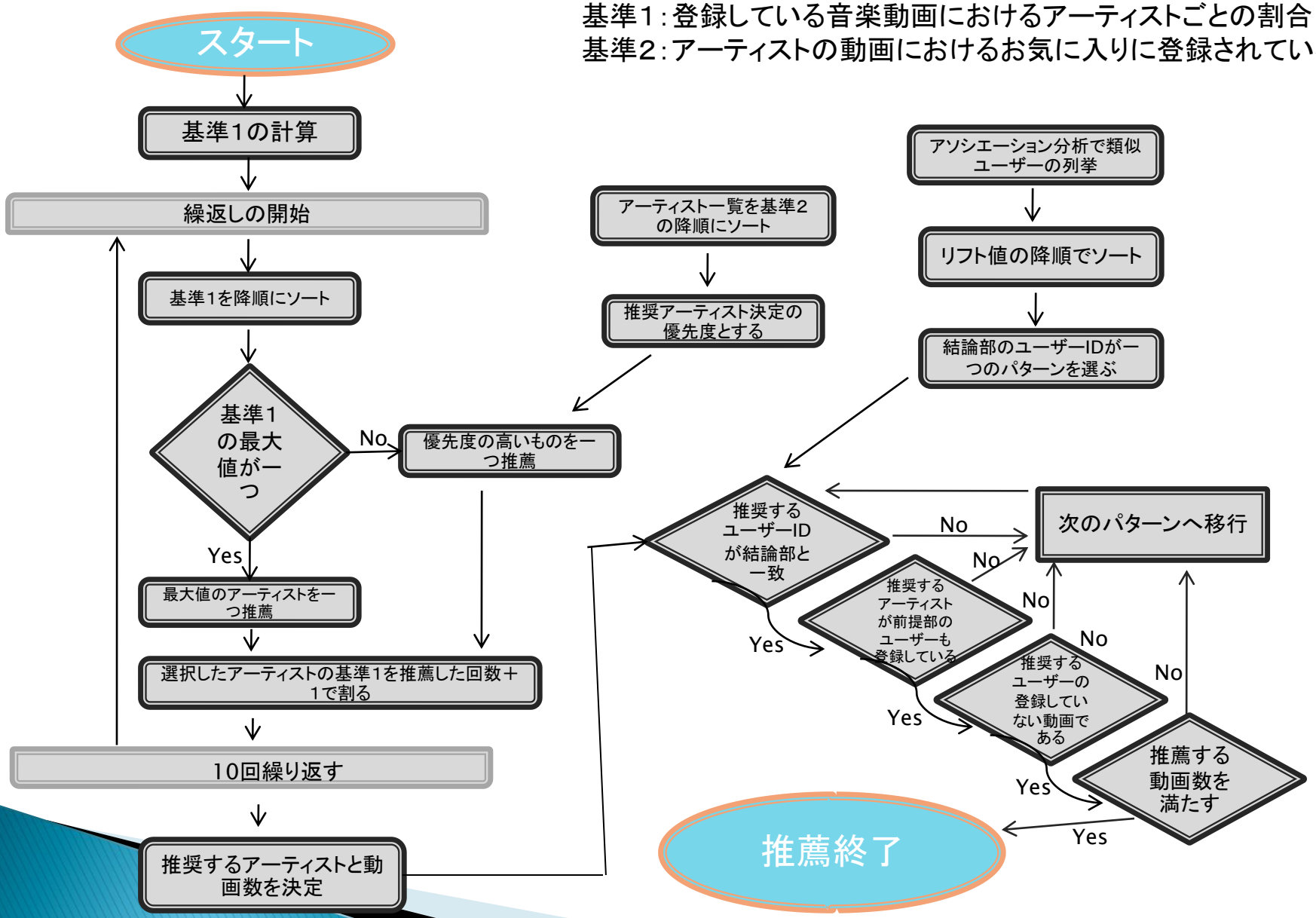
前提	結論	信頼度	サポート	Lift	Conviction	ルール.数	前提.数	結論.数	キー.数
uid-5361+uid-1961	uid-1708	100	0.622	9.137	-1	12	12	211	1928
uid-7+uid-1909+uid-5361	uid-1708	100	0.622	9.137	-1	12	12	211	1928
uid-6439+uid-1	uid-1708	100	0.571	9.137	-1	11	11	211	1928
uid-6439+uid-405	uid-1708	100	0.571	9.137	-1	11	11	211	1928
uid-421+uid-746	uid-1708	100	0.519	9.137	-1	10	10	211	1928
uid-405+uid-38	uid-1708	100	0.519	9.137	-1	10	10	211	1928
uid-5361+uid-1244	uid-1708	100	0.519	9.137	-1	10	10	211	1928
uid-421+uid-1244	uid-1708	100	0.519	9.137	-1	10	10	211	1928
uid-6439+uid-24	uid-1708	100	0.519	9.137	-1	10	10	211	1928
uid-5361+uid-6439	uid-1708	94.444	0.882	8.63	16.03	17	18	211	1928
uid-10011+uid-38	uid-1708	92.857	0.674	8.485	12.468	13	14	211	1928
uid-3174+uid-38	uid-1708	92.857	0.674	8.485	12.468	13	14	211	1928
uid-1909+uid-5361	uid-1708	92.857	0.674	8.485	12.468	13	14	211	1928
uid-5361+uid-405	uid-1708	92.308	0.622	8.435	11.577	12	13	211	1928
uid-1708+uid-1909+uid-5361	uid-7	92.308	0.622	10.786	11.887	12	13	165	1928
uid-7+uid-421+uid-5361	uid-1708	92.308	0.622	8.435	11.577	12	13	211	1928
uid-1708+uid-10011+uid-1909	uid-7	92.308	0.622	10.786	11.887	12	13	165	1928
uid-7+uid-10011+uid-1909	uid-1708	92.308	0.622	8.435	11.577	12	13	211	1928
uid-1909+uid-54	uid-1708	91.667	0.571	8.376	10.687	11	12	211	1928
uid-3174+uid-1973+uid-2718	uid-405	91.667	0.571	21.04	11.477	11	12	84	1928

⋮

この結果含め759パターンを列挙

動画推奨のフロー

基準1：登録している音楽動画におけるアーティストごとの割合
基準2：アーティストの動画におけるお気に入りに登録されている割合



推奨する動画の決定手順

1. ・アソシエーション分析の結果出てきたパターンの前提部と結論部にあるユーザーIDを一つのグループと考え、リフト値で降順に並べた際にユーザーIDを上から検索し、IDのあったグループをユーザーとの類似グループとする。
2. ・推奨する動画はグループ内にいる他のユーザーがお気に入りに登録している同アーティストの動画かつ推奨するユーザーが登録していないものを選択する。
3. ・所属するグループに推奨する個数以上に動画があった場合、ユーザー全体でお気に入りに登録されている数が多い順に選択する。
4. ・所属するグループに推奨するアーティストの動画が含まれない場合は、次にリフト値の高いグループに移り同様に推奨アーティストの動画を検索する。
5. ・どのグループにも所属していない、もしくはどのユーザーが所属するグループから動画が見つからなかった場合は推奨するアーティストの動画をお気に入り登録数の降順に並べ、上から必要数選択する。

方法の検定

- お気に入り登録20個未満のユーザーのうち19個を登録しているユーザーからランダムに9個を抜き出した後に同様の手順で動画を9つ選択した場合いくつ当てることができるかで検証する。

- 今回は19個登録しているユーザーからランダムで5人を選び、提案した手法で動画の推薦を行ってみた。

結果

- ランダムで抜き出した9件×5人、合計45件のうち抜き出した動画を当てたのは2件であった。

動画マスターのデータからランダムに選ぶよりは確率は高くなっているが、当たっていると自信を持って言える数値ではない。

改善の余地あり!!

今後の展望

- 今回は音楽系の動画のみに絞っての推奨であったが、TMSのことばネットワークで見られる分類のように、他のジャンルで同様の推奨方法が可能なのか、それともまったく別の方法を提案するか吟味する必要がある。
- また、今回一件でも当てたユーザーとそうでなかったユーザーの違いを見つけ、今後の分析に利用していきたい。