Twitter上のあるユーザの意外な情報抽出手法の提案

甲南大学　知能情報学部
◎大原 啓詳　灘本 明代

背景

Twitter

140文字以内の短文の投稿

文章が短い

様々なデバイスで利用可能

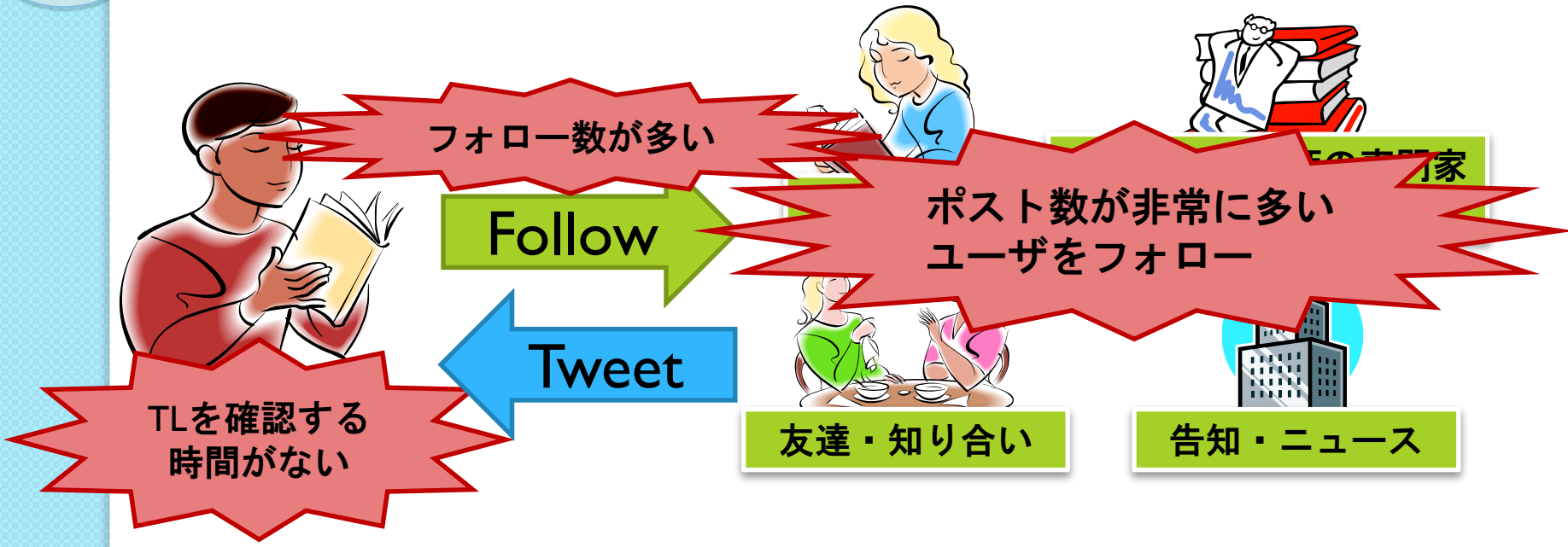
リアルタイム情報発信

多くのがユーザが気軽に投稿

テーマに囚われない多種多様な投稿

背景

Twitterにおけるユーザ同士の繋がり



ツイートの投稿と閲覧

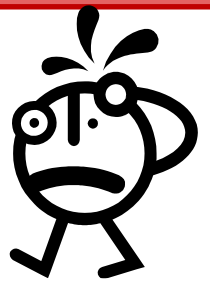
背景

タイムライン上のツイート

時間がない
フォローが多い

全て確認

困難



効率化が必要

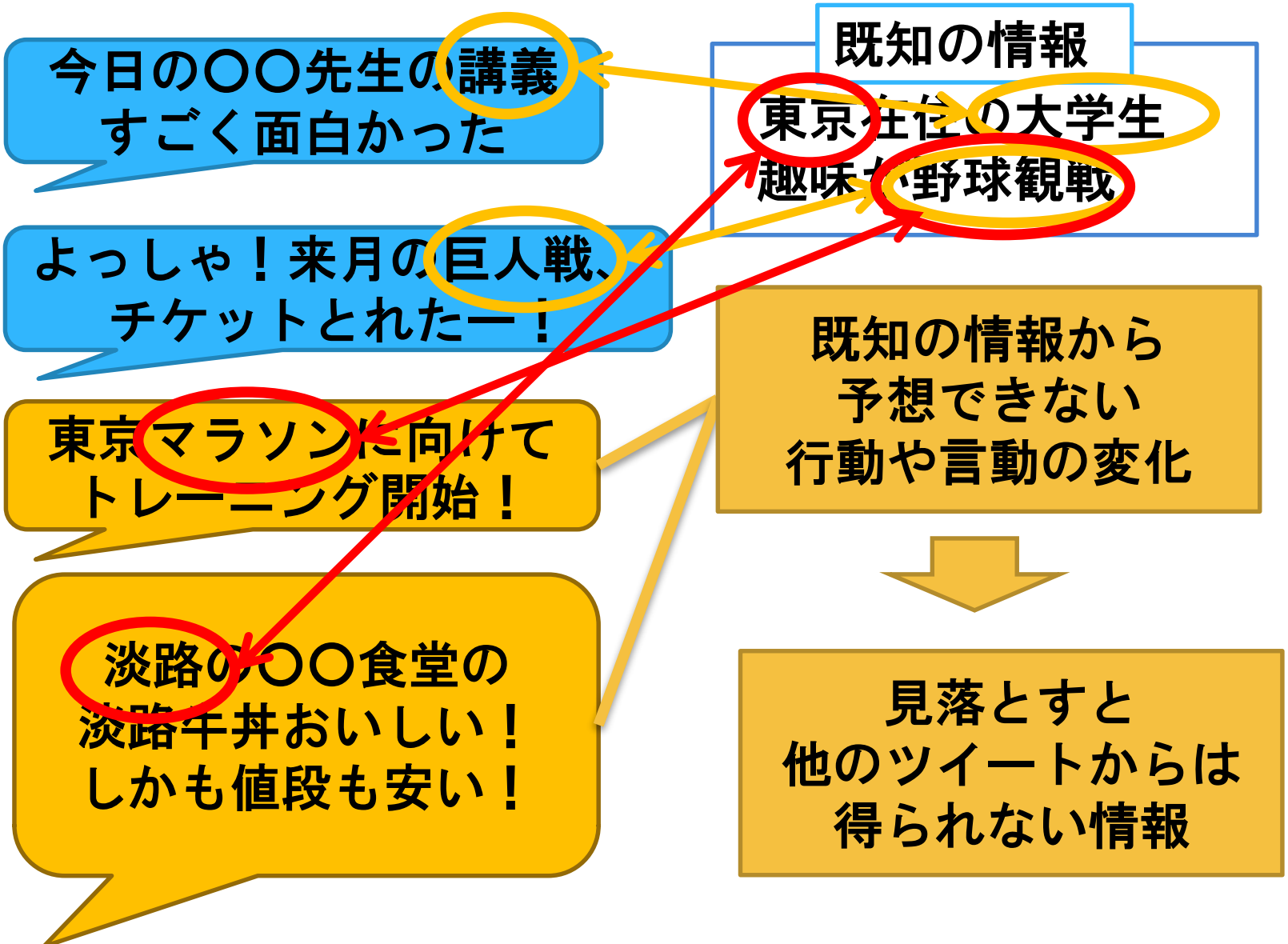
閲覧者が欲しい情報

- 変化のあった情報
- 新鮮味のある情報

フォローしているあるユーザの…

予想外の行動・言動

背景


今日の〇〇先生の講義
すごく面白かった
よっしゃ！来月の巨人戦、
チケットとれたー！
東京マラソンに向けて
トレーニング開始！
淡路の〇〇食堂の
淡路牛丼おいしい！
しかも値段も安い！
既知の情報
東京在住の大学生
趣味が野球観戦
既知の情報から
予想できない
行動や言動の変化
見落とすと
他のツイートからは
得られない情報

目的

既知の情報から
予測できない
行動や言動の変化

→ **意外な情報**

情報（ツイート）の発信者の
意外な情報

↓ 抽出

有益・貴重なツイートの発見

意外な情報の分類

アトリビュート意外情報

ツイートの持つ投稿時間や頻度
位置情報に関する意外な情報

コンテンツ意外情報

ツイートの文そのものに含まれる
感情や行動に関する意外な情報

アトリビュート意外情報

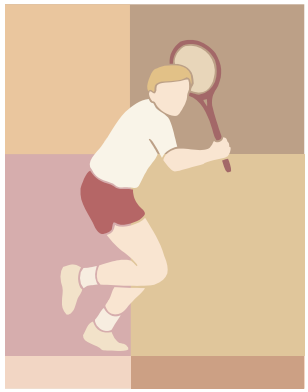
アトリビュート意外情報

- Tweetの投稿時間
- Retweet数
- Favorite数
- 位置情報

コンテンツ意外情報

コンテンツ意外情報

ツイートの文に含まれる
- 位置変化の分かる地名・イベント名
- 感情や気分の変化
- 普段は行わない行動
- 興味・嗜好の変化



アトリビュート意外情報の抽出手法

アトリビュート意外情報

- Tweetの投稿時間
- 位置情報

完全に
データのみから
判断可能

- Retweet数
- Favorite数

内容についても
考慮する必要がある

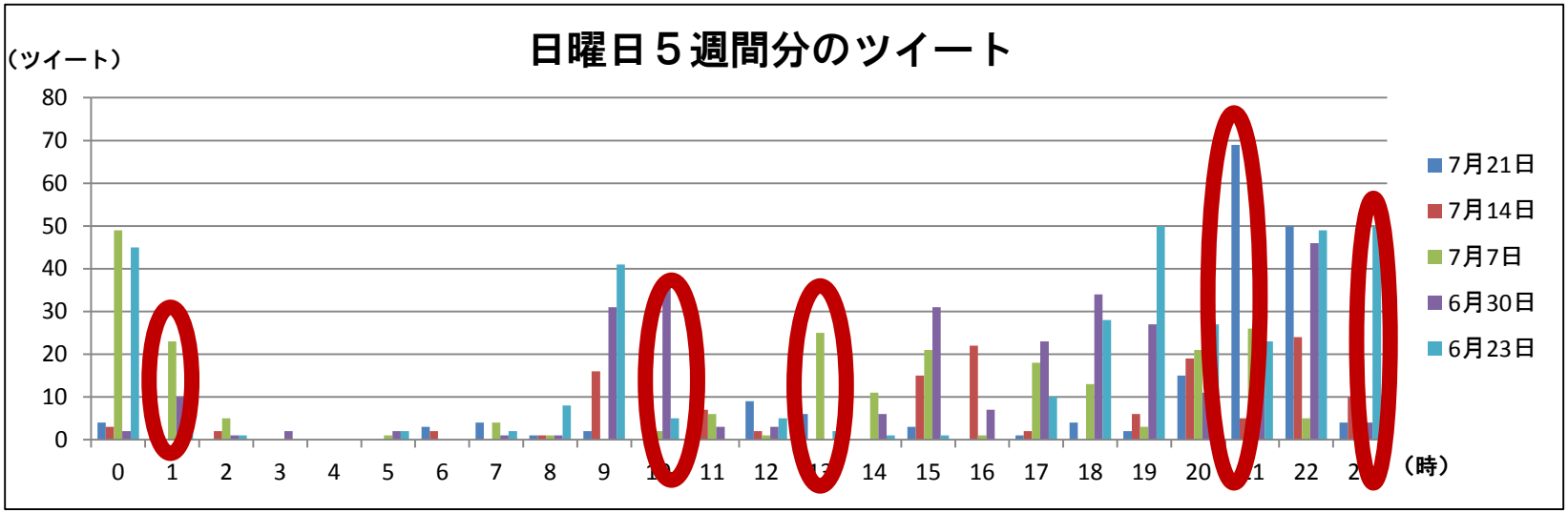
ツイートの分析　ツイート数

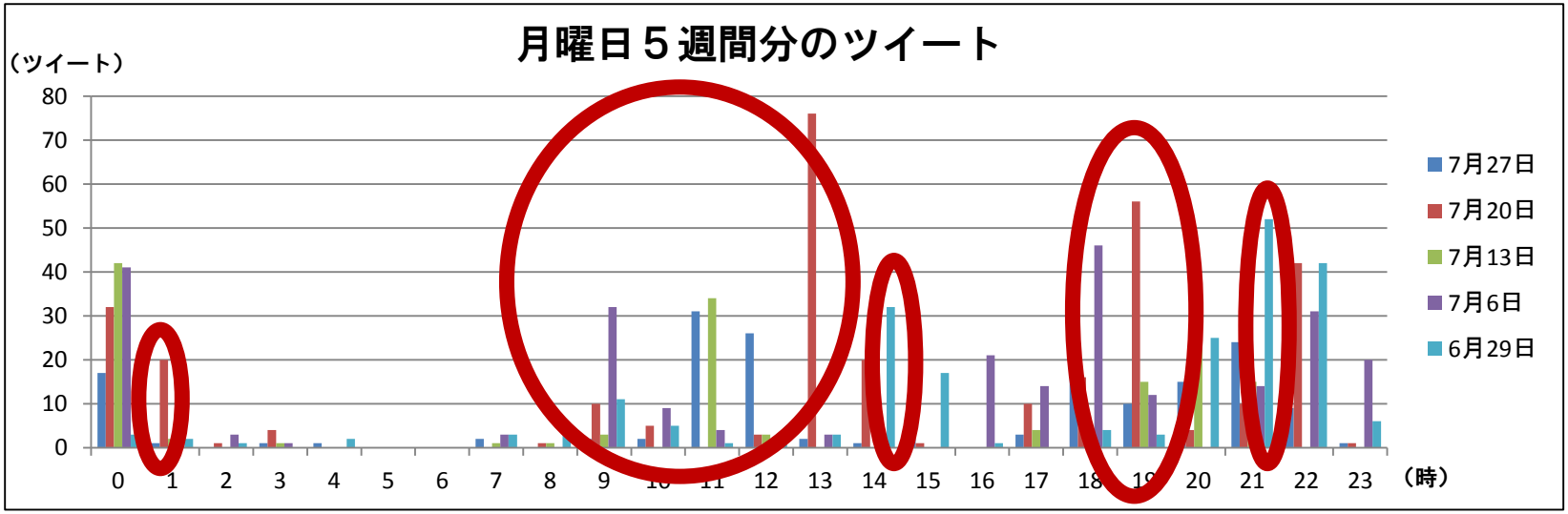
1時間ごとのツイート数をグラフ化

対象データ：ユーザ5名の5週間分のツイート
計21161ツイート

目的：ユーザごと、曜日ごとの
ツイートの投稿パターンの規則性
が存在するかの調査

ツイートの分析　ツイート数





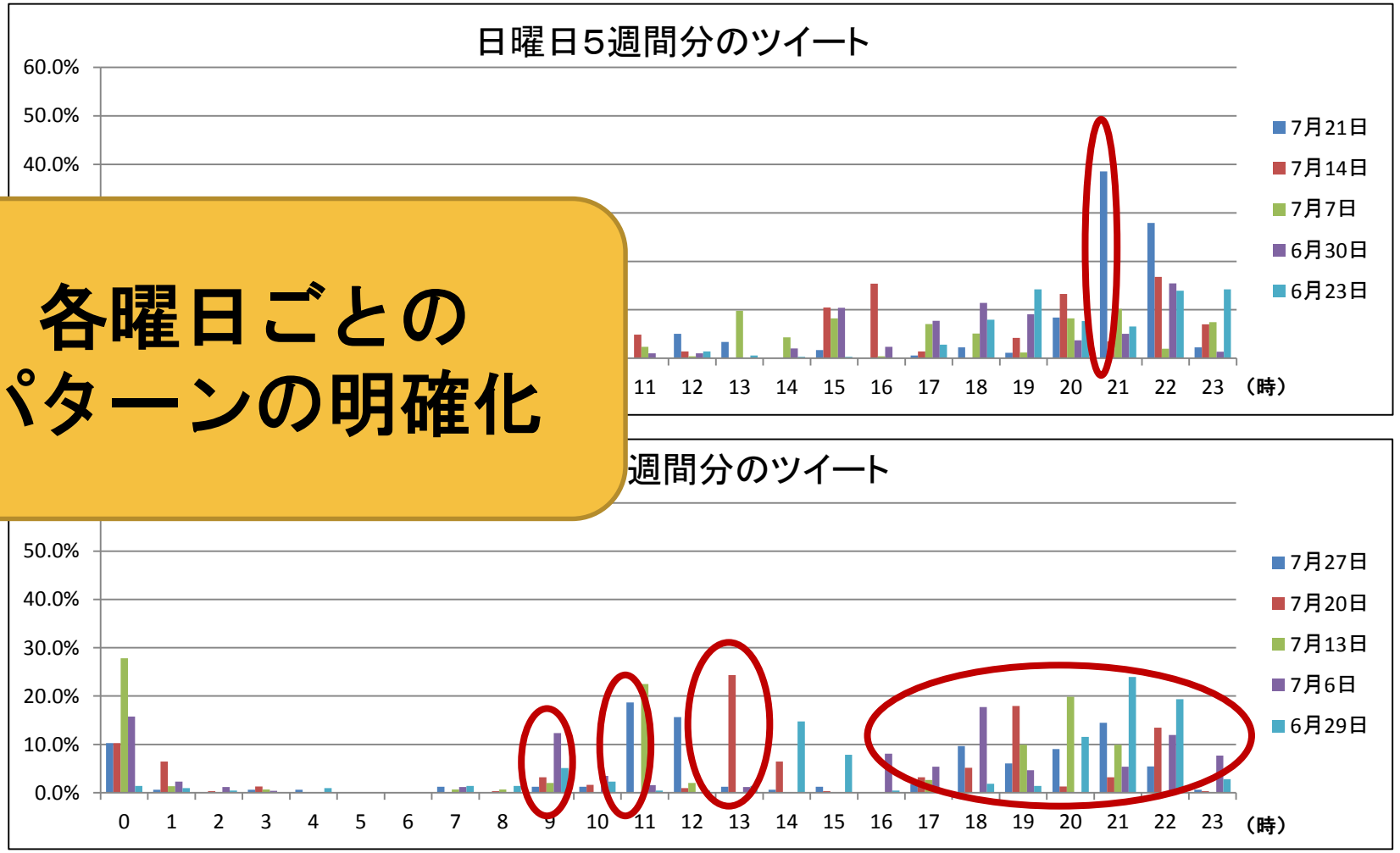
ツイートの分析　ツイート割合

ユーザのツイートはある程度の規則性を持つ
例：昼間のツイート数が平日は多い
　　深夜のツイート数が土日は多い

同じ曜日でもツイート数にはばらつきがある

曜日ごとのパターンの分析
⇒一日における各時間でのツイート割合

ツイートの分析　ツイート割合



各曜日ごとの
パターンの明確化

コンテンツ意外情報の抽出手法

コンテンツ意外情報

- 位置変化

アトリビュート意外情報と合わせて考慮

- 感情や気分
- 行動
- 興味・嗜好

判断の為膨大なツイートを読む必要がある

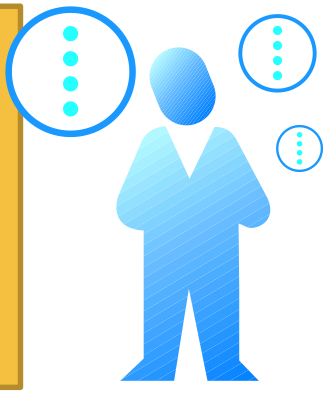
ツイートのみで判断

コンテンツ意外情報の抽出手法

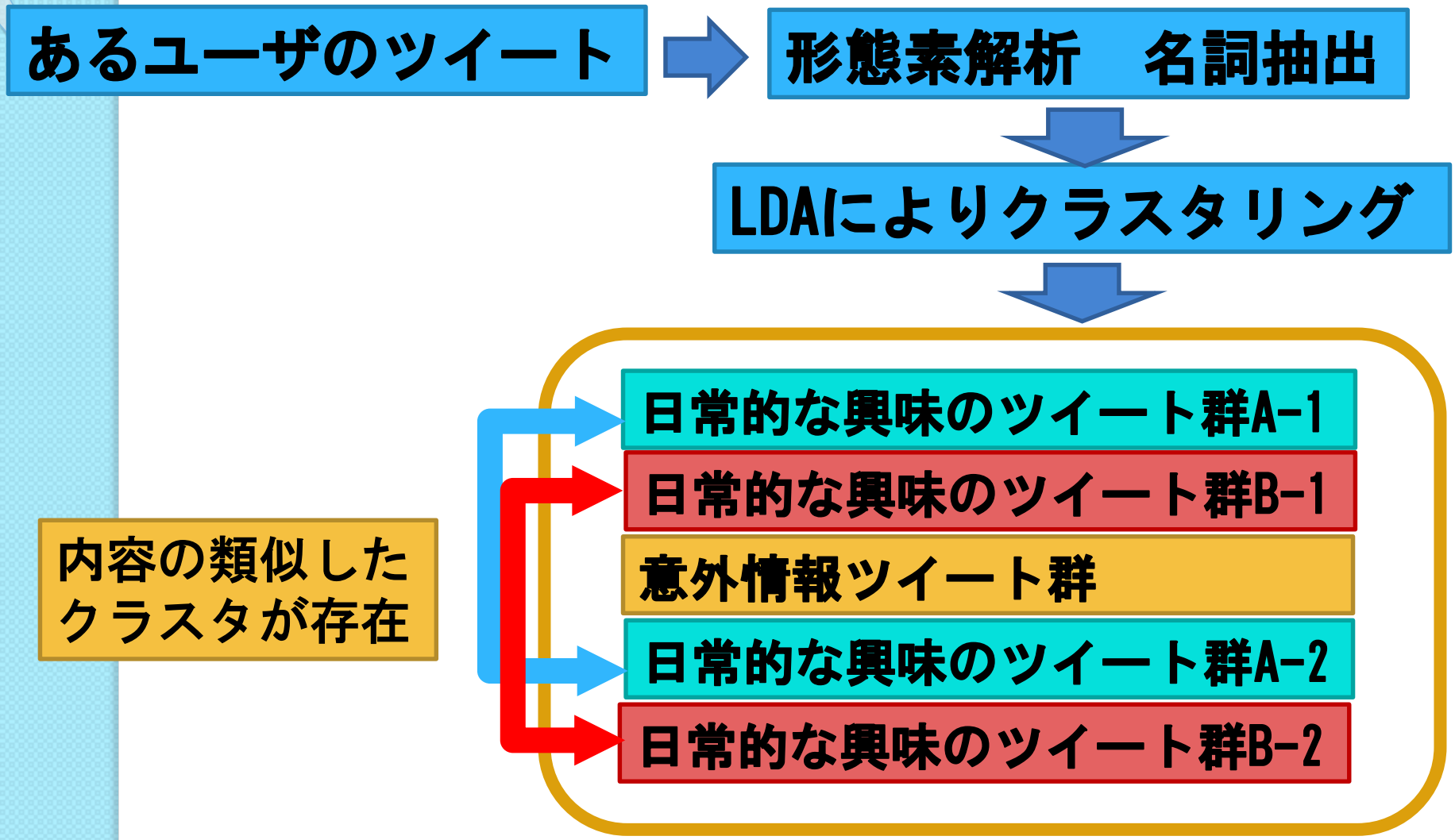
行動・興味に
関する意外情報

普段のツイートとは
異なる話題に関するツイート

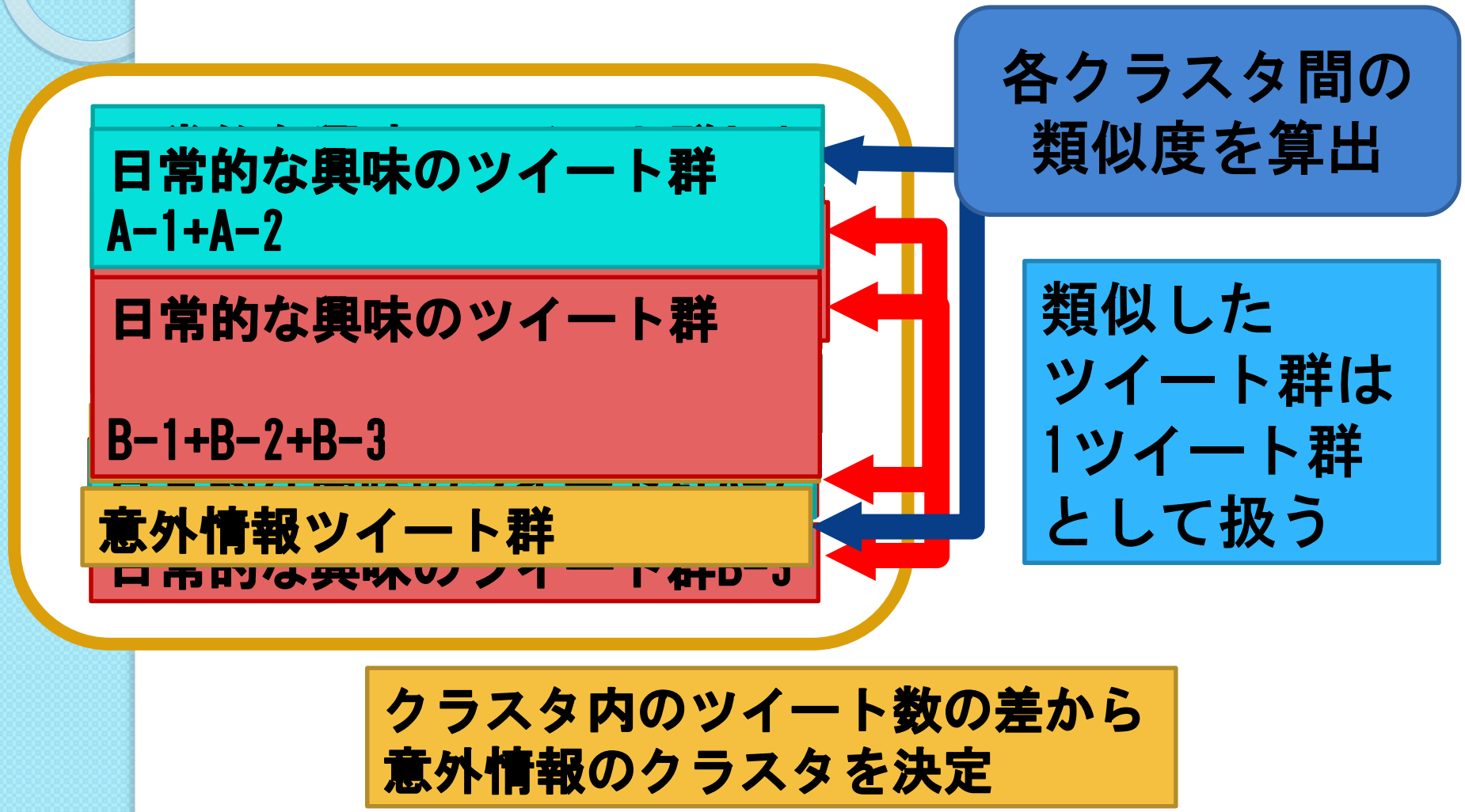
ツイートを話題ごとに
クラスタリングし
意外な話題に関する
ツイートのクラスタを抽出



Tweetのクラスタリングによる意外情報ツイートの抽出



Tweetのクラスタリングによる意外情報ツイートの抽出



コンテンツ意外情報の抽出手法

- 評価実験

実験データ

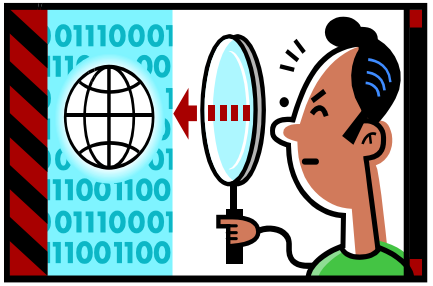
- 実験対象アカウント　3アカウント
- ツイート数　各2000ツイート（計6000ツイート）

正解データ

- 意外情報の定義に基づき人手で判断

各パラメータ（予備実験により決定）

- LDAによる分類　20クラスタ
- cos類似度の閾値　0.3以上
- 意外情報クラスタとする基準
  含まれるツイート数が100ツイート以下



コンテンツ意外情報の抽出手法

- 実験結果

ユーザ	意外情報クラスタ数	意外情報ツイート数	適合率	再現率	F値
A	3	199	0.18593	0.11315	0.140684
B	2	167	0.08982	0.060241	0.072115
C	1	49	0.163265	0.041522	0.066207

- 考察

適合率・再現率、共に低い

・適合率が低い原因
⇒ツールからの定型ツイートのクラスタの発生

・再現率が低い原因
⇒LDAによるクラスタリングの失敗
⇒根本的な原因はツイートの短さと考えられる

まとめと今後の課題

まとめ

- ユーザの意外情報の定義と分類を行った
- アトリビュート意外情報抽出のための分析
- コンテンツ意外情報抽出のための
  システムの作成と評価実験

課題

- 大規模なツイートの分析
- 分析データに基づくアトリビュート意外
  情報の抽出手法の具体的なロジック構成
- コンテンツ意外情報抽出システムの
  為のクラスタリング手法や
  クラスタの類似比較手法の再検討
- 複合的な意外情報の抽出手法の考案

ツイートの分析

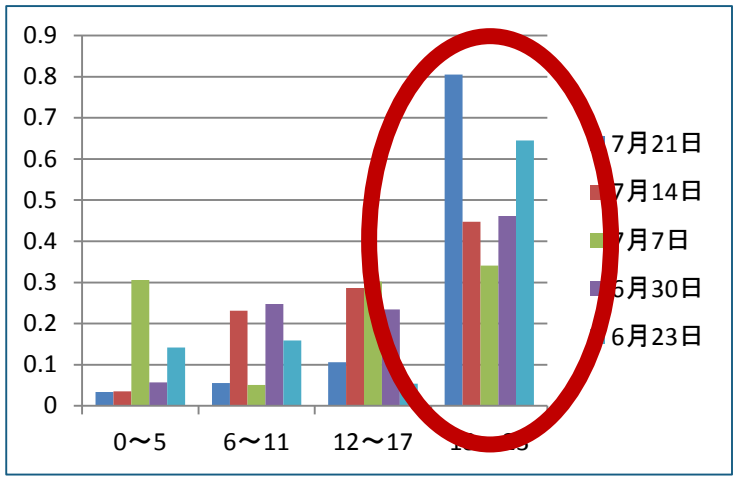
投稿パターン分析のための時間の分割方法

候補

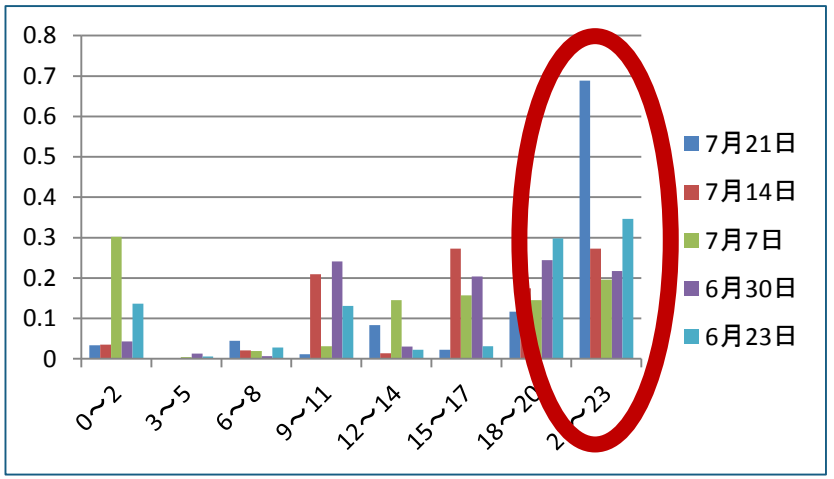
- **(a) 6時間ごと4等分**
- **(b) 3時間ごと8等分（気象庁準拠）**
- **(c) 1時間ごと24等分**

各区間時間ごとの1日に占めるツイート割合について、グラフの比較

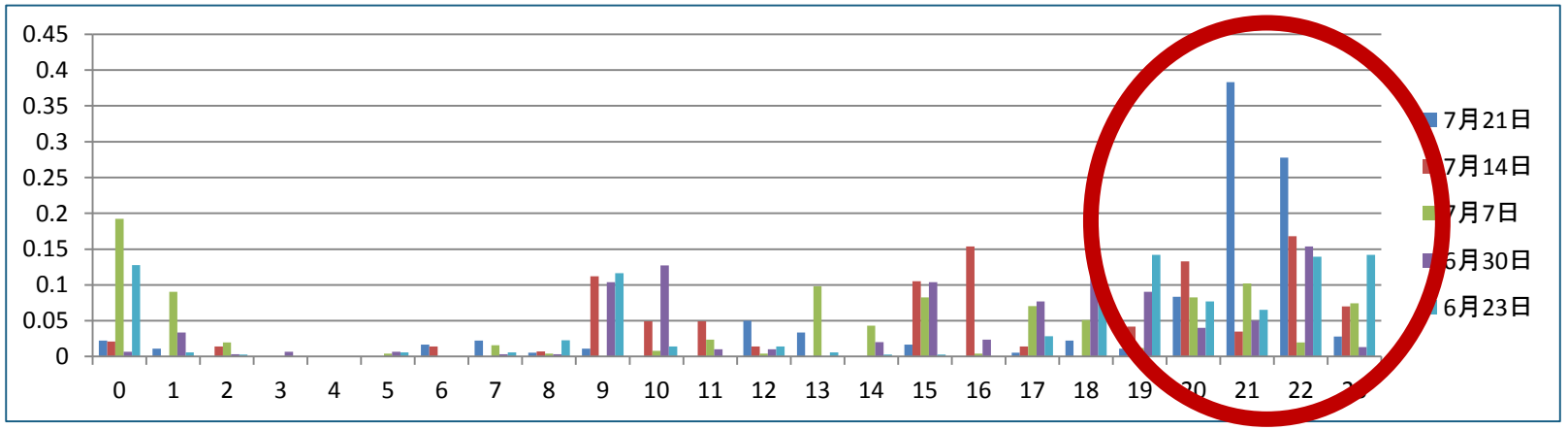
ツイートの分析　時間区分



(a)6時間ごと



(b)3時間ごと



(c)1時間ごと

コンテンツ意外情報の抽出手法

・適合率を著しく下げた例

愛のある質問をして下さい。 http～
理想形の質問をして下さい。 http～
落ち着いて質問して下さい。http～

△△音楽室光の結晶
△△音楽室未来への帰り道
△△音楽室 ザギンデビュー
△△音楽室億万笑者