

AWSを利用した Twitterデータ収集のプロトタイプ

2012/08/30

クラウド・テクノロジー研究部会 メンバー

橋本 健介

TwitterAPIを使用することで、ツイートの内容や、ツイートしている人のユーザ情報等の情報を収集することができます。

Twitterの情報を集める事で分析できる事例

・世界のその日の「気分」

→ 金融業界では、ツイートからその日の気分を読み取り、世界が幸せそうときには買い、不幸そうときには売る

・会員登録したユーザの関心事

→ 会員登録したユーザのツイートを分析して、ユーザが関心するような情報を推薦

- どのような情報を獲得できるか
- 獲得の手順と制限事項
- 実際の動作

- **どういった情報を獲得できるか**
- 獲得の手順と制限事項
- 実際の動作

```
"StatusJSONImpl{createdAt=Sat Jul 28 20:21:49 JST 2012, id=229174821559365632, text='いい位置でフリーキックを得た。行け #aitc3', source='<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>', isTruncated=false, inReplyToStatusId=-1, inReplyToUserId=-1, isFavorited=false, inReplyToScreenName='null', geoLocation=null, place=null, retweetCount=0, wasRetweetedByMe=false, contributors=null, annotations=null, retweetedStatus=null, userMentionEntities=[], urlEntities=[], hashtagEntities=[HashtagEntityJSONImpl{start=18, end=24, text='aitc3'}], user=UserJSONImpl{id=78009134, name='小林茂', screenName='skoba84', location='東京都世田谷区', description='', isContributorsEnabled=false, profileImageUrl='http://a0.twimg.com/profile_images/440817270/___-___normal.jpg', profileImageUrlHttps='https://si0.twimg.com/profile_images/440817270/___-___normal.jpg', url='null', isProtected=false, followersCount=57, status=null, profileBackgroundColor='9AE4E8', profileTextColor='333333', profileLinkColor='0084B4', profileSidebarFillColor='DDFFCC', profileSidebarBorderColor='BDDCAD', profileUseBackgroundImage=true, showAllInlineMedia=false, friendsCount=62, createdAt=Mon Sep 28 22:09:08 JST 2009, favouritesCount=772, utcOffset=32400, timeZone='Tokyo', profileBackgroundImageUrl='http://a0.twimg.com/images/themes/theme16/bg.gif', profileBackgroundImageUrlHttps='https://si0.twimg.com/images/themes/theme16/bg.gif', profileBackgroundTiled=false, lang='ja', statusesCount=1025, isGeoEnabled=true, isVerified=false, translator=false, listedCount=4, isFollowRequestSent=false}}"
```

Status

作成日時

ステイタスのID

このステイタスがfavoriteされているか否か

このステイタスがTruncateされているか否か

このステイタスがリツイートされたものだった場合、元のステイタス

このステータスをツイートしたユーザーのUserオブジェクト

このステイタスの内容

このステイタスをツイートしたクライアント名 or Web

緯度/経度を格納するGeoLocationオブジェクト

Status

作成日時

ステイタスのID

このステイタスがfavoriteされているか否か

このステイタスがTruncateされているか否か

このステイタスがリツイートされたものだった場合、元のステイタス

このステータスをツイートしたユーザーのUserオブジェクト

このステイタスの内容

このステイタスをツイートしたクライアント名 or Web

緯度/経度を格納するGeoLocationオブジェクト

user

favoriteしている数

followerの数

followingの数

UserId

Geotaggingを有効にしているかどうか

ツイートを非公開にしているかどうか

認証済みアカウントかどうか

現在地

名前

ユーザー名

タイムゾーン

その他のURL

協定世界時との差分

Status

作成日時

ステイタスのID

このステイタスがfavoriteされているか否か

このステイタスがTruncateされているか否か

このステイタスがリツイートされたものだった場合、元のステイタス

このステータスをツイートしたユーザーの**Userオブジェクト**

このステイタスの内容

このステイタスをツイートしたクライアント名 or Web

緯度/経度を格納する**GeoLocationオブジェクト**

赤色・・・ビッグデータの元にできそう

user

favoriteしている数

followerの数

followingの数

UserId

Geotaggingを有効にしてるかどう

ツイートを非公開にしてるかどう

認証済みアカウントかどうか

現在地

名前

ユーザー名

タイムゾーン

その他のURL

協定世界時との差分

赤色・・・ビッグデータの元にできそう

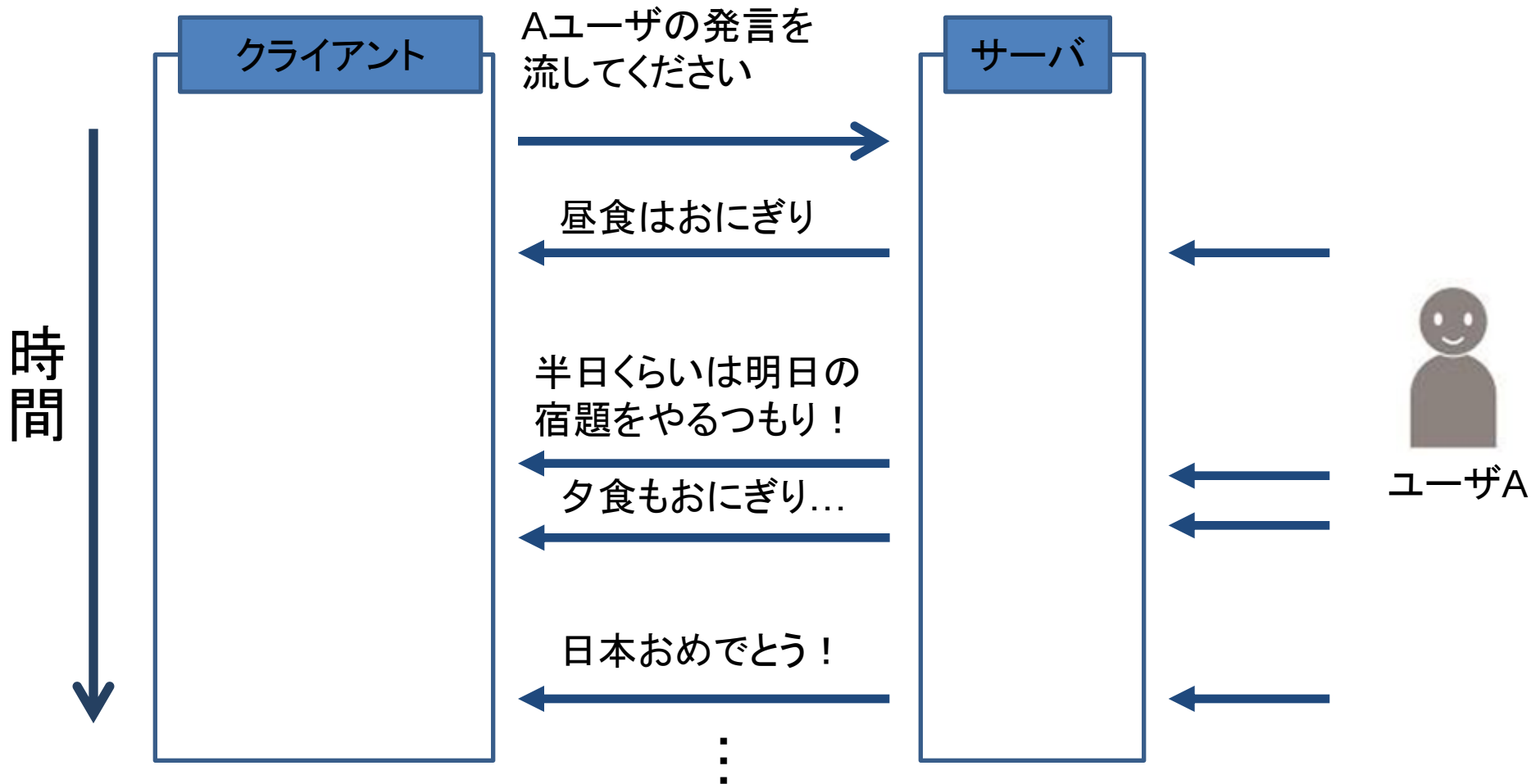
- **どういった情報を獲得できるか**
- **獲得の手順と制限事項**
- **実際の動作**

- どういった情報を獲得できるか
- 獲得の手順と制限事項
- 実際の動作

- ・ストリーム型
条件に従ってサーバから
データが流れてくる
- ・問い合わせ型
獲得したい情報を指定して
毎回問い合わせる

- ストリーム型
条件に従ってサーバから
データが流れてくる
- 問い合わせ型
獲得したい情報を指定して
毎回問い合わせる

ストリーム型



ストリーム型

- Streamed Tweets
... 一般に流れているツイートを条件付きで取得
- ユーザーストリーム
... ユーザーのタイムライン、@関連、DMなどを取得
- サイトストリーム
... 多数のユーザーのタイムライン、@関連、DMなどを取得

ストリーム型

- Streamed Tweets
... 一般に流れているツイートを条件付きで取得
- ユーザーストリーム
... ユーザーのタイムライン、@関連、DMなどを取得
- サイトストリーム
... 多数のユーザーのタイムライン、@関連、DMなどを取得

Streamed Tweets

... 一般に流れているツイートを条件付きで取得

- sample ... パブリックなツイートから約1%のツイート
- firehose ... すべてのパブリックなツイート
- links ... リンクを含むツイート
- retweet ... リツイート
- filter ... 指定した条件にマッチするツイート
 - ユーザID
 - ロケーション
 - キーワード

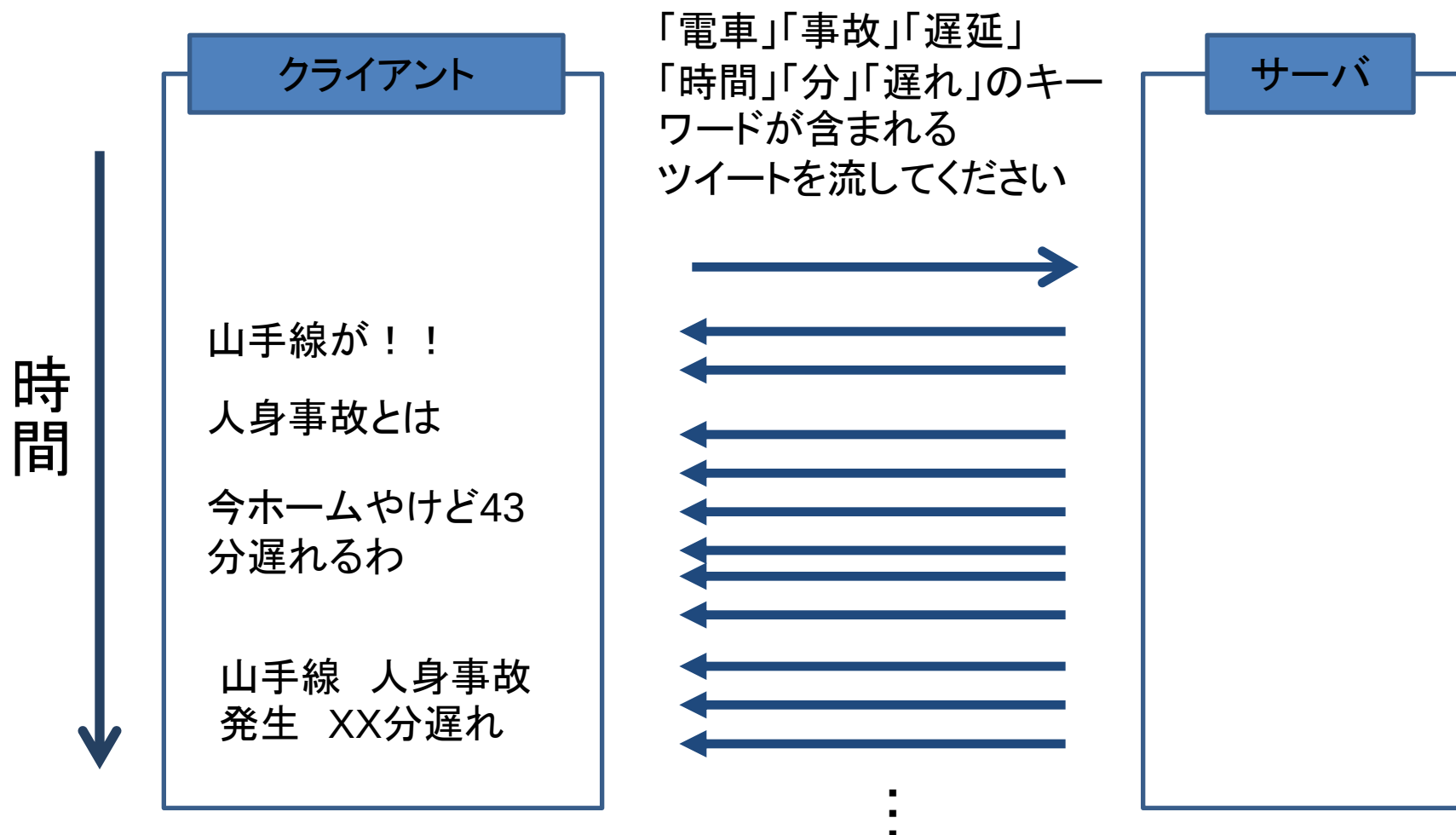
電車の遅延に関するツイートを集めて
分析することで、日本中の電車遅延
MAPとか遅延速報みたいなサービス
を作れないかな？



Streamed Tweets

... パブリックなツイートを取得

- sample ... パブリックなツイートから約1%のツイート
- firehose ... すべてのパブリックなツイート
- links ... リンクを含むツイート
- retweet ... リツイート
- **filter** ... 指定した条件にマッチするツイート
 - ユーザID
 - ロケーション
 - **キーワード**



- ストリーム型
条件に従ってサーバからデータが流れてくる



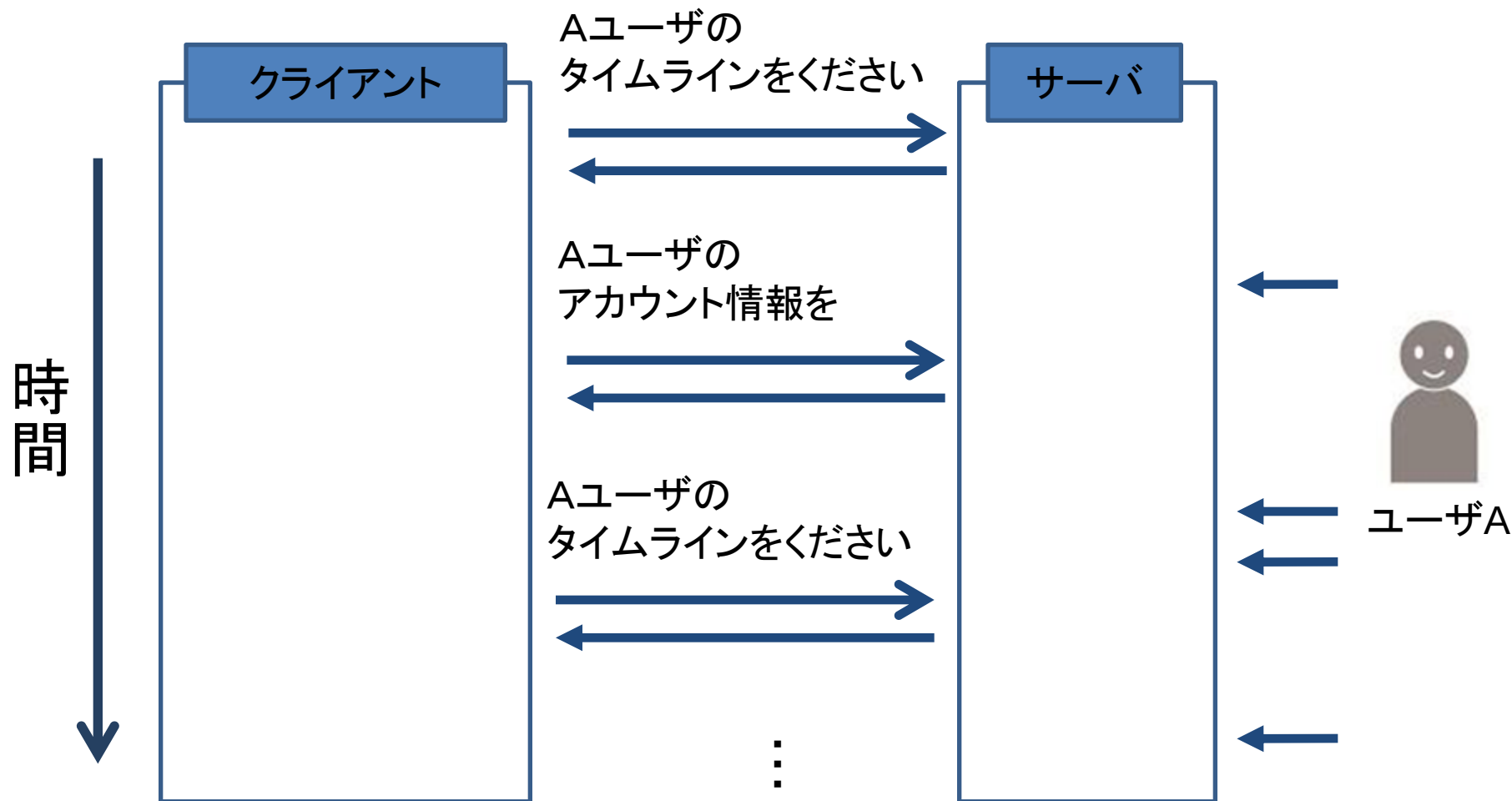
制限 / 注意事項

- ・ **ほとんどの企業は全てのツイートを全獲得できない**
 - firehoseを使えるのは、今のところ大手7企業のみ
sampleは誰でも使えるが、全体の1%程度
(平日11:30～12:10で85909 ツイート取得)
 - **大量のツイートを無作為に獲得するには、
複数アカウントでsampleをいくつも動かし、
重複を排除していくなどしなければならない**
- ・ **リクエストしてからのツイートしか獲得できない**
 - 過去のツイートは取り出せない
 - **過去のツイートを獲得するには
問い合わせ型を使わなければならない**

- ストリーム型
条件に従ってサーバから
データが流れてくる
- 問い合わせ型
獲得したい情報を指定して
毎回問い合わせる

- ストリーム型
条件に従ってサーバから
データが流れてくる
- 問い合わせ型
獲得したい情報を指定して
毎回問い合わせる

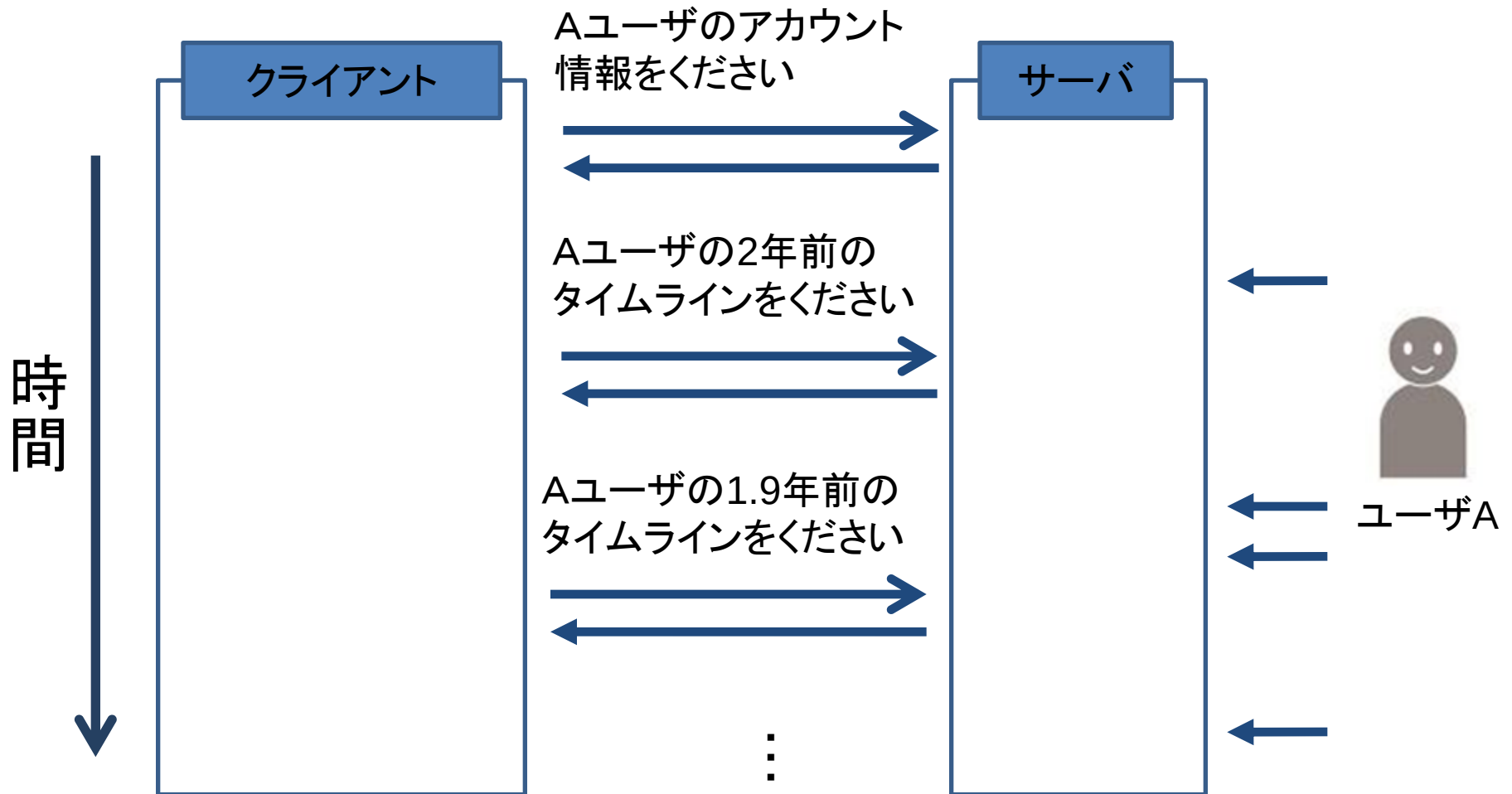
問い合わせ型



会員登録してくれたユーザの関心事
を過去までさかのぼって分析したい



問い合わせ型



- 問い合わせ型

獲得したい情報をそのつど問い合わせる



制限 / 注意事項

- 1時間あたりのリクエスト回数は350回
 - さらに、オリンピックの競技中等、
負荷が高い時は、回数の制限が変動する
 - 残り回数を意識したアルゴリズム
- ツイートにロケーション情報を含ませる事を許可している人はほとんどいない
 - 位置情報を絡めた分析をしたい場合、
キーワードから推測する等、ひと工夫が必要

- ストリーム型
条件に従ってサーバから
データが流れてくる
- 問い合わせ型
獲得したい情報を指定して
毎回問い合わせる

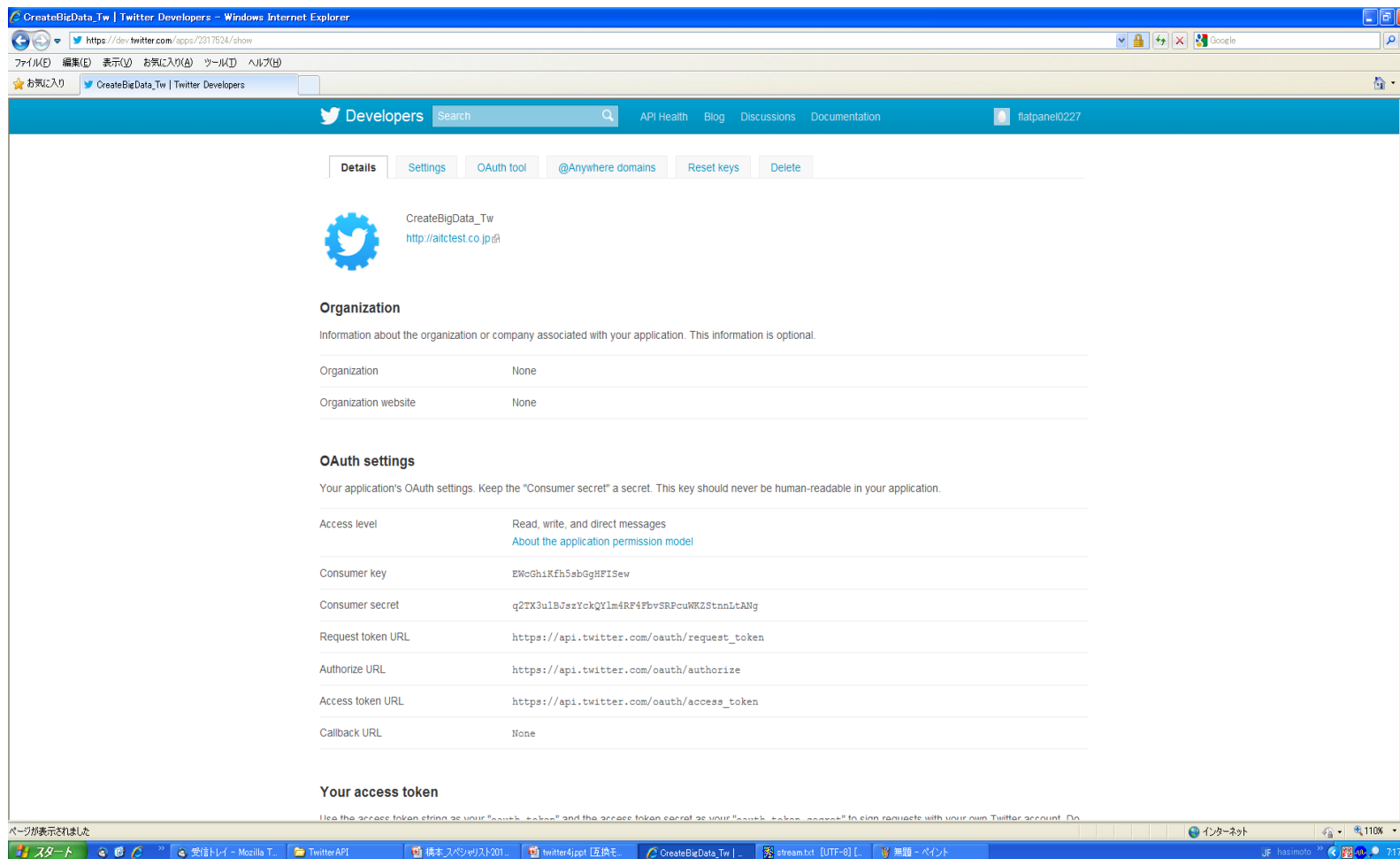
動作手順

ユーザを介したAuth認証

具体的には以下が必要

- Consumerkey
- Cnsumersecret
- Accesstoken
- Accessecret

Consumerkey と Cnsumersecret は毎回同じでいいが、
Accesstoken と Accessecret はユーザ毎に必要



1. Twitter Developer にアプリ登録
2. Consumerkey と Cnsumersecret 獲得

3. Consumerkey と Cnsumersecret で
リクエストトークン発行
4. リクエストトークン使ってtwitter.comに
ユーザにアプリ連携許可するためのURLを
発行してもらう



5. ユーザがtwitterアカウントでログイン、アプリ連携承認
6. twitter.comが発行するPinCodeとリクエストトークンで Accesstoken と Accesssecret 獲得

- ユーザ操作

twitterアカウントでのログインとアプリ連携の認証

- 開発者

アプリ登録と、各ユーザのアクセストークンの管理
ユーザのTwitterパスワード等はあずからなくていい

- どういった情報を獲得できるか
- 獲得の手順と制限事項
- 実際の動作

- どういった情報を獲得できるか
- 獲得の手順と制限事項
- 実際の動作

- ・先端ITコンソーシアムの会員を対象に、
#aitcがついたツイートを収集
- ・サーバはAWSのEC2
- ・DBサーバはAWSのDynamoDB

実際の動作

DynamoDB Management Console - Windows Internet Explorer

https://console.aws.amazon.com/dynamodb/home?explorername=tweetregion=ap-northeast-1

ファイル(F) 編集(E) 表示(V) お気に入り(I) ツール(T) ヘルプ(H)

お気に入り DynamoDB Management Console

Services Edit Shortcut Kensuke Help

DynamoDB: Explore Table: tweet

Region: Asia Pacific (Tokyo)

List Tables Browse Items Put Item Item Details

Scan Get Go 1 to 10 of 20 loaded items

status	num	time
"StatusJSONImpl{createdAt=Sun Aug 05 21:19:01 JST 2012, id=232088316160385024, text='40km位の地点。まだ先頭は4人の集団。日本の選手の情報は全く分からない、かなり遅れている模様。 #aitc3', source='Teewee', isTruncated=false, inReplyToStatusId=-1, inReplyToUserId=-1, isFavorited=false, inReplyToScreenName='null', geoLocation=null, place=null, retweetCount=0, wasRetweetedByMe=false, contributors=null, annotations=null, retweetedStatus=null, userMentionEntities=[], urlEntities=[], hashtagEntities=[HashtagEntityJSONImpl{start=49, end=55, text='aitc3'}], user=UserJSONImpl{id=78009134, name='小林茂', screenName='skoba84', location='東京都世田谷区', description='', isContributorsEnabled=false, profileImageUrl='http://a0.twimg.com/profile_images/440817270/____normal.jpg', profileImageUrls='https://s10.twimg.com/profile_images/440817270/____normal.jpg', url='null', isProtected=false, followersCount=57, status=null, profileBackgroundColor='9AE4E8', profileTextColor='333333', profileLinkColor='0084B4', profileSidebarFillColor='DDFFCC', profileSidebarBorderColor='BDDCAD', profileUseBackgroundImage=true, showAllInlineMedia=false, friendsCount=62, createdAt=Mon Sep 28 22:09:08 JST 2009, favouritesCount=783, utcOffset=32400, timeZone='Tokyo', profileBackgroundImageUrl='http://a0.twimg.com/images/themes/theme16/bg.gif', profileBackgroundImageUrls='https://s10.twimg.com/images/themes/theme16/bg.gif', profileBackgroundTiled=false, lang='ja', statusesCount=1074, isGeoEnabled=true, isVerified=false, translator=false, listedCount=4, isFollowRequestSent=false}}"	"230"	"2012.8.5.21.19.1"
"StatusJSONImpl{createdAt=Thu Aug 02 06:44:41 JST 2012, id=230781120541765632, text='体操 男子個人総合で、内村航平が金メダルに輝いた。やったね。おめでとう。そしてありがとう。団体戦では調子悪かったみたいだが、個人戦では完璧のようだ。あの跳馬で16.266を採ったのはさすが、やっぱり内村航平は強い。 #aitc3', source='Teewee', isTruncated=false, inReplyToStatusId=-1, inReplyToUserId=-1, isFavorited=false, inReplyToScreenName='null', geoLocation=null, place=null, retweetCount=0, wasRetweetedByMe=false, contributors=null, annotations=null, retweetedStatus=null, userMentionEntities=[], urlEntities=[], hashtagEntities=[HashtagEntityJSONImpl{start=108, end=114, text='aitc3'}], user=UserJSONImpl{id=78009134, name='小林茂', screenName='skoba84',		

© 2008 - 2012, Amazon Web Services LLC or its affiliates. All rights reserved. | Feedback | Support | Privacy Policy | Terms of Use | An amazon.com company

ページが表示されました

スタート 受信トレイ - Mozilla T... TwitterAPI 橋本、スペシャリスト201... twitter4j.ppt [互換モ... DynamoDB Managem... stream.txt [UTF-8] [... 無題 - バイント

インターネット 110% 7:35

- どのような情報を獲得できるか
- 獲得の手順と制限事項
- 実際の動作

TwitterAPIを使ってデータ収集



さまざまな分析を行って、
今までにないサービスを考えて
いきましょう