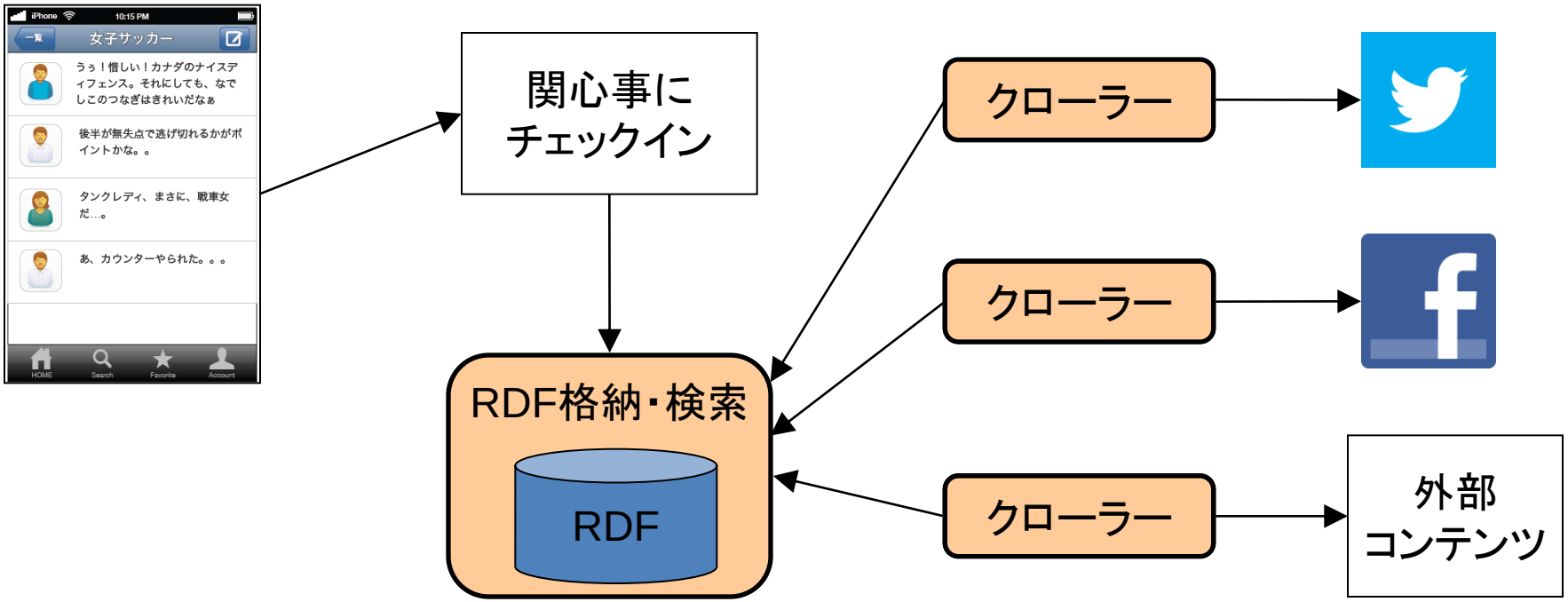
AITC
先端ＩＴ活用推進コンソーシアム

Project LA
バックグラウンド処理と課題

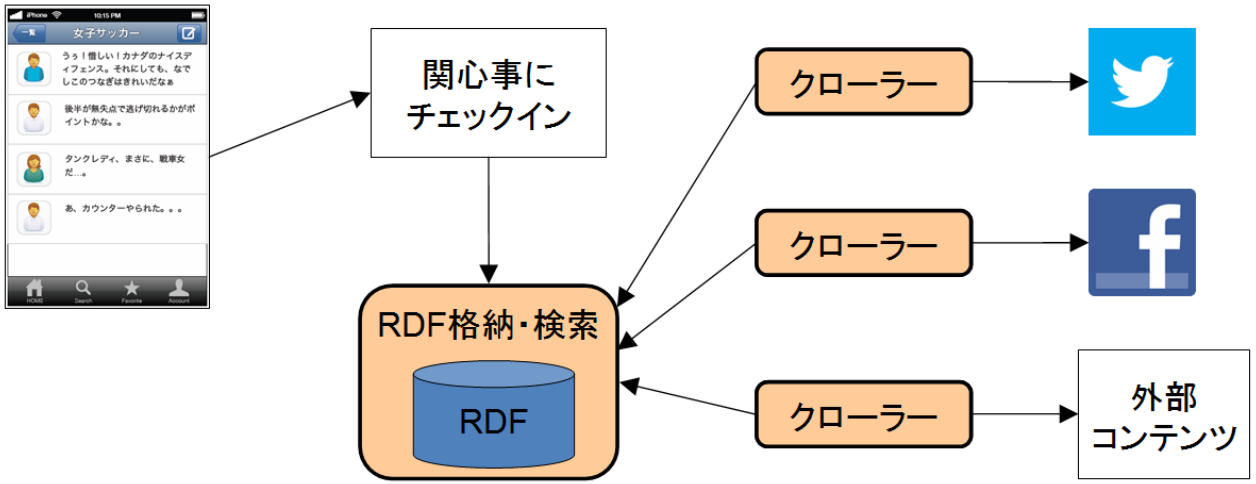
バックグラウンド処理

- RDFを格納・検索するサービス
- SNSのクローラー
- 外部コンテンツのクローラー



技術的な課題

- 大量データ
  - 対象がSNSと外部コンテンツなので、大量に発生
  - GUIから呼び出すので、リアルタイムの応答性能が必要
- セキュリティ
  - 利用者ごとに認可をもらわないと取れない部分もある
  - どこまでをクローリング対象にするのか
- データ構造
  - 違うところから集めたものを1つのRDFの中に入れる



RDFを格納・検索するサービス

AITC
先端ＩＴ活用推進コンソーシアム

- Internet上で試験運転中
  - 登録は、RDFをもらってそのまま登録
  - 検索は、SPARQLのクエリをそのまま実行
  - まだデータは少ないが、レスポンスはまずまず
  - 抽象化してAPIにする予定
- 技術的課題
  - 大量データの格納
    - RDFをKey Value Storeに格納してみる
    - SPARQLの実行性能の測定
  - 構造化された大量データを使ってみる
    - 機械学習の元データとして使ってみたい

SNSのクローラー

- SNSをクローリングし、RDF化して格納
- 技術的課題
  - 各SNSの認可方式の違い、APIの違い
    - ProjectLAに対して1回の認可で済むように
  - 構造の違いをRDFにどう反映するか
    - Twitterの返信、Facebookのコメント
  - 認可していない人の返信・コメントをどうするか？
    - 収集してここに貯めてもいいのか？
    - 「返信の返信」「返信の返信の返信」どこまで？

外部コンテンツのクローラー

- 外部コンテンツをクローリング、RDF化して格納
- 技術的課題
  - 外部の様々なデータ構造への対応
    - NewsML、気象庁防災情報XML、CSV、など
  - データの表現方法の違いへの対応
    - 住所と緯度経度、など
  - 外部コンテンツの更新を検出
- どの外部コンテンツを収集するのか？
  - まだ具体的なものは決まっていません

RDFサーバ

AITC 先端ＩＴ活用推進コンソーシアム

- 様々な情報を１つのRDFに格納
  - SNS
    - 発言、返信・コメント
  - 外部コンテンツ
    - 気象庁防災情報XML、その他
- 串刺しでクエリーを実行
  - SPARQLでこのようなクエリーを書きたい
    - 「台風が来ている時は、どんな関心事が多いか」
    - 「晴れの日と雨の日の、発言内容の違い」