筑波大学大学院博士課程

システム情報工学研究科特定課題研究報告書

Hadoop を用いた大規模ログデータに対する相関ルールマイニングシステムの開発 ―データ加工システムの開発と成果物の取りまとめ―

永田佑輔

(コンピュータサイエンス専攻)

指導教員　田中二郎

２０１２年　３月

概要

近年、社会の高度情報化と情報発信の低コスト化により、日々、大量のデータが生成されている。また、記録媒体の大容量化と通信の高速化により、膨大なデータの蓄積や流通が可能になった。そのため、企業が保有するデータ量が急激に増加している。

ビジネス環境の変化や計算機性能の向上により、膨大なデータの有効活用する試みが行われている。例えば、各種セキュリティ基準のコンプライアンスチェック、株価や交通システム利用状況などの各種分野において、大規模ログデータの利用が進んでいる。

大規模なログデータから、知識を得る方法として、データマイニングがある。データマイニングを利用することで、データの中から相関ルールのような規則性や特定のパターンを得ることができる。また、ログデータとは、毎日の気温や視聴率、商品販売数など、時間軸で連続しているデータのことで、これらデータを解析することは、未来予測や市場調査を行う上で重要である。そして、データ中のパターンの同時生起に注目した相関ルール分析の有効性が知られている。しかし、一般的に、大規模なデータの分析処理には時間がかかる。

大規模ログデータの分析処理を高速化する方法として、並列処理がある。ログデータを時間軸や空間軸を基準に分割し、処理を高速化している。また、大規模ログデータを高速に処理できる基盤として､クラウドや大規模分散処理フレームワーク Hadoop が注目されている。

そこで、我々は大規模分散処理フレームワーク Hadoop を用いて、大規模ログに対する相関ルールマイニングシステムを開発する。また、多様なデータ形式に対応するために、プラグインを用いて、入力データの中から処理に必要なデータを抽出する。

筆者は、このプロジェクトで、データ加工システムのプラグイン機能とヒストグラム作成機能、データ加工機能の開発を担当した。また、データ分析システムの「ある場所に関する物質ごとの相関ルール」を抽出する機能と抽出した相関ルールを場所ごとに表示する KML ファイルの作成機能の開発も担当した。

プラグイン機能とは、多様なデータ形式に対応するために、入力データの中から処理に必要なデータを抽出するものである。プラグインに入力データの属性を保持し、入力データとプラグインのマッチングをとり、合致する入力データのみを処理対象とする。ヒストグラム作成機能とは、利用者に表示するヒストグラムを作成する。物質ごとに度数を集計する。データ加工機能とは、入力データを中間データのフォーマットに加工する。「ある場所に関する物質ごとの相関ルール」を抽出する機能とは、同一の場所において、異なる物質間の相関ルールを抽出するものである。抽出した相関ルールを場所ごとに表示する。KML ファイルの作成機能とは、抽出した相関ルールを Google Earth で表示するために、相関ルールを KML ファイル形式で出力する。

成果物のとりまとめでは、成果物の作成予定日数と実績を比較した。その結果、作成予定日数に対して､実績に大きな遅れが出ていることがわかった｡筆者はその理由を四つ考えた。一つ目は、事前調査(5～8 月)に時間がかかったからだと考える。二つ目は議事録にもれがあり、次回までのアクション項目が行えなかったからだと考える。三つ目はドキュメントにあいまいな表現と定性的な表現が多く、会議を行うごとに指摘を受けていたからだと考える。四つ目はレビューを行う会議の直前にドキュメントを送付していたからだと考える。

目次

図目次

表目次

第1章　はじめに

1.1　　プロジェクトの概要

近年、社会の高度情報化と情報発信の低コスト化により、日々、大量のデータが生成されている。また、記録媒体の大容量化と通信の高速化により、膨大なデータの蓄積や流通が可能になった。そのため、企業が保有するデータ量が急激に増加している。

ビジネス環境の変化や計算機性能の向上により、膨大なデータの有効活用する試みが行われている。例えば、各種セキュリティ基準のコンプライアンスチェック、株価や交通システム利用状況などの各種分野において、大規模ログデータの利用が進んでいる。

大規模なログデータから、知識を得る方法として、データマイニングがある[1]。データマイニングを利用することで、データの中から相関ルールのような規則性や特定のパターンを得ることができる。また、ログデータとは、毎日の気温や視聴率、商品販売数など、時間軸で連続しているデータのことで、これらデータを解析することは、未来予測や市場調査を行う上で重要である。そして、データ中のパターンの同時生起に注目した相関ルール分析の有効性が知られている[2]。

しかし、一般的に、大規模なデータの分析処理には時間がかかる。例えば、クックパッドでは消費者の潜在的な食材へのニーズを求めるために、ユーザが入力した膨大な検索ログのキーワード解析を月別／地域別に行っており、その処理に 7000 時間かかると報告されている[3]。

大規模ログデータの分析処理を高速化する方法として、並列処理がある。ログデータを時間軸や空間軸を基準に分割し、処理を高速化している。また、大規模ログデータを高速に処理できる基盤として、クラウドや大規模分散処理フレームワーク Hadoop が注目されている。先のクックパッドの事例では、Hadoop を導入することで、処理時間を 7000 時間から 30 時間に短縮できたとも報告されている。

そこで、我々は大規模分散処理フレームワーク Hadoop を用いて、大規模ログに対する相関ルールマイニングシステムを開発する。また、多様なデータ形式に対応するために、プラグインを用いて、入力データの中から処理に必要なデータを抽出する。

1.2　　扱うログデータ

ログデータには、大気汚染の測定値のように、空間情報と時間情報を含むものがある。大気環境の分析者にとって、測定値と空間情報、時間情報を関連付けた分析結果は有益であり、環境評価／予測等に利用される。

そこで、我々は一例として本システムで扱うログデータを大気汚染ログデータとする。そして、大気汚染の測定値から、汚染物質に関する相関ルールを抽出するシステムの開発を行う。

1.3　本報告書の構成

本報告書は全 8 章から構成される。

第 2 章では、本プロジェクトの前提となる知識を述べる。第 3 章では、開発するシステムの機能要件と想定する利用者、システム構成等を述べる。第 4 章では、筆者が担当した機能の開発について述べる。第 5 章では、本システムを、実行速度と抽出ルールの観点から評価した結果を述べる。第 6 章では、本プロジェクトの開発体制と開発スケジュール、開発の推移について述べる。第 7 章では、成果物のとりまとめについて述べる。第 8 章では、本プロジェクトの成果から出した結論について述べる。

第2章　前提知識

2.1　Hadoop

Hadoop は主に Yahoo! Inc.の Doug Cutting 氏によって開発が進められているオープンソースソフトウェアである。Google の基盤ソフトウェアである Google File System[4]と、MapReduce[5]のオープンソース実装となっている。また、HDFS（Hadoop Distributed File System）[6]、Hadoop MapReduce Framework から構成されている。

Hadoop はすべて Java で記述されており、MapReduce 処理を書く場合も基本的には Java でプログラムを書くことが想定されている。ただし、Hadoop Streaming[7]という拡張パッケージを用いると、C/C++・Ruby・Python など任意の言語と標準入出力を用いて MapReduce 処理を書くことも出来る。

現在、Hadoop は分散コンピューティングに関連するサブプロジェクトの集合体である(図 2-1)。これらのプロジェクトは、Apache Software Foundation によってホストされている[8]。

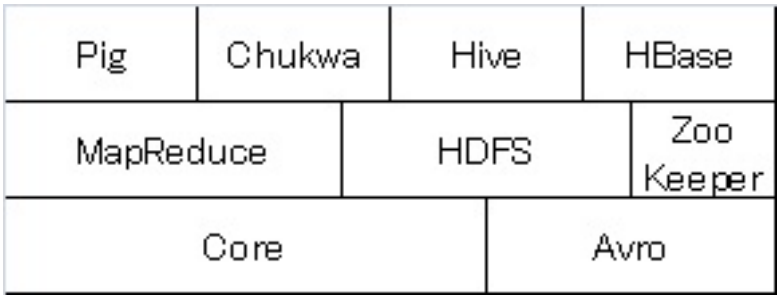


図 2-1　Hadoop のサブプロジェクト

本プロジェクトでは MapReduce 処理と HDFS を用いて、分散処理を行う。次に、MapReduce について述べる。図 2-2 に MapReduce 処理の流れを示す。Map 処理は解析前のデータを入力として、入力データから必要な情報を抜き出して Reduce 処理に渡す。Reduce 処理は、Map 処理の出力を入力とし、必要な計算を行い、出力する。

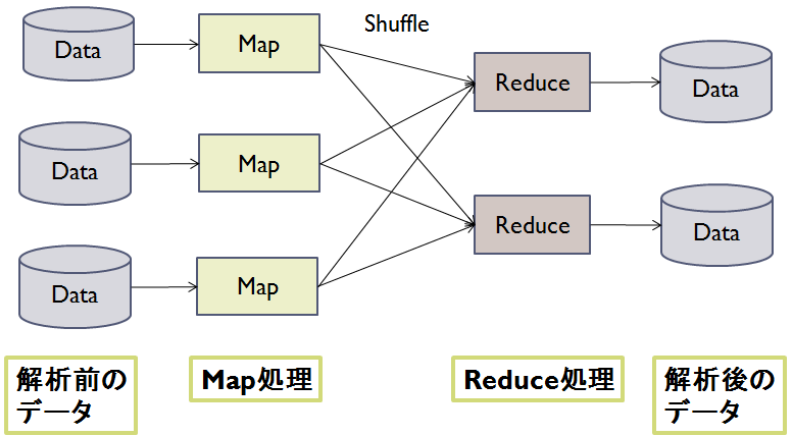


図 2-2　MapReduce 処理

次に HDFS について述べる。HDFS は Hadoop 分散ファイルシステムであり、非常に大きなファイルを保存するために設計されたファイルシステムである。コモディティハードウェアによって構成されるクラスタで動作する。

図 2-3 に HDFS のアーキテクチャを示す[9]。HDFS は NameNode と DataNode という 2 つのサーバで構成されている。HDFS クライアントはこれら 2 つのサーバと通信し、ファイル操作を行う。

DataNode は実際のデータを保持するサーバであり、データをブロックという固定サイズの単位に分割し、保持している。NameNode はファイルシステムのメタデータ（ディレクトリ構造やファイルのアクセス権など）を管理するサーバである。

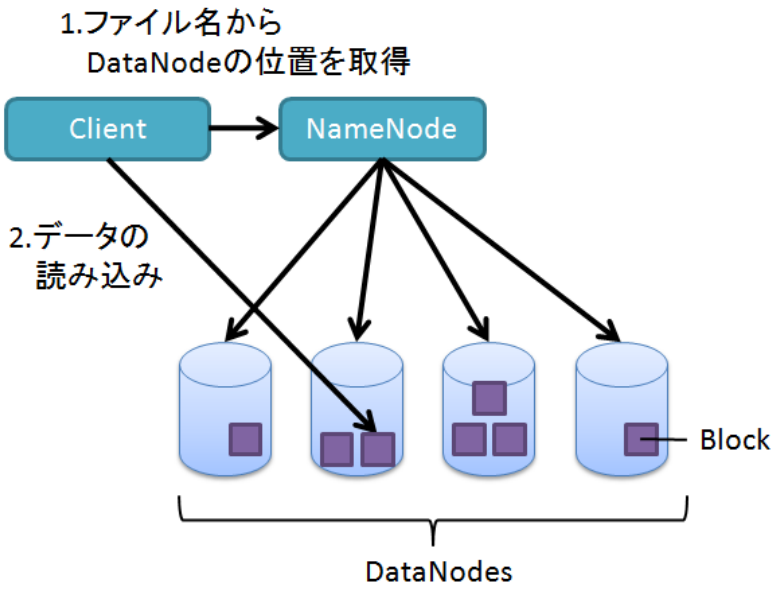


図 2-3　HDFS アーキテクチャ

2.2 相関ルールマイニング

2.2.1 相関ルール

マーケットで売られている個々の商品をアイテム、一人の顧客が購買した商品のリストをトランザクションと呼ぶ。全ての顧客のトランザクションを解析すると、例えば、「バターを買った顧客は、その 80%がパンと牛乳も買っており、この 3 種の商品すべてを買った人は全顧客の 4%である。」というような知識が得られる。これを次のように表したものが相関ルールである[10]。

条件部		結論部	支持度	確信度
[バター]	⇒	［パン、牛乳］; supp=4%、conf=80%		

ここで、ルールの条件部、結論部ともに複数のアイテムを含む場合がある。また、ルール中の全てのアイテムが現れるようなトランザクションの割合を支持度(supp)、条件部のアイテムを購買した顧客の中で結論部のアイテムを買った人の割合を確信度(conf)と呼ぶ。

相関ルールを抽出する手法として、相関ルール分析があり、本プロジェクトではその中のアプリオリアルゴリズムを用いる。

2.2.2 アプリオリアルゴリズム

アプリオリアルゴリズム[11]について述べる。まず、アルゴリズムを簡潔に示すために以下の記号を定義する。

D:トランザクションのデータベース
δ:最小支持度数
$F[I_s](D,\delta)$:データベースの中の最小支持度数を
満たすアイテムの集合
C_k:kアイテムの候補集合
L_k:kアイテムの頻出アイテム集合

次にアルゴリズムについて述べる。

$\mathrm{Input}: D,\delta$

$\mathrm{Output}: \bigcup_{i=1}^{k} L_k$

$L_1 = \{F[i](D,\delta)\}$
$\mathrm{for}(k=2; L_k \neq 0; k++)\mathrm{do}$
  $C_k = \mathrm{apriori\_gen}(L_{k-1})$
  for all $t \in T$ do
    $C_{kt} = \mathrm{subset}(C_k, t)$
  end
  comput$|C_{kt}|$ kアイテム候補の支持数を計算
  $L_k = \{F[C_{kt}](D,\delta)\}$
end

$C_k = \text{apriori\_gen}(L_{k-1})$: K−1アイテム頻出集合L_{k-1}から
kアイテム候補集合C_kを生成
$C_{kt} = \text{subset}(C_k, t)$: C_kとトランザクションの照合で頻出の
kアイテムの候補を生成

2.2.3 相関ルールの評価指標

抽出した相関ルールの評価指標[12]として、支持度(support)、確信度(confidence)、リフト値(lift)がある。

支持度とは、アイテム集合 X と Y が同時に起こる確率である。X と Y を含むトランザクション数 σ(X∪Y)を全体のトランザクション数 M で割った値で表される。

$$\text{supp}(X \Rightarrow Y) = \frac{\delta(X \cup Y)}{M}$$

確信度とは、アイテム集合 X が起こったという条件の下で、アイテム集合 Y が起こる確率である。X と Y を含むトランザクション数 σ(X∪Y)を、条件 X を含むトランザクション数 σ(X)で割った値で表される。

$$\text{conf}(X \Rightarrow Y) = \frac{\delta(X \cup Y)}{\delta(X)} = \frac{\text{supp}(X \Rightarrow Y)}{\text{supp}(X)}$$

リフトとは、抽出したルールの重要性を表した値である。確信度を supp(Y)で割った値で表される。

$$\text{lift}(X \Rightarrow Y) = \frac{\text{conf}(X \Rightarrow Y)}{\text{supp}(Y)}$$

第3章　相関ルールマイニングシステムの開発

3.1　　システム全体像

本システムは、利用者に入力データから抽出した相関ルールを提供する。図 3-1 にシステムの全体像を示す。

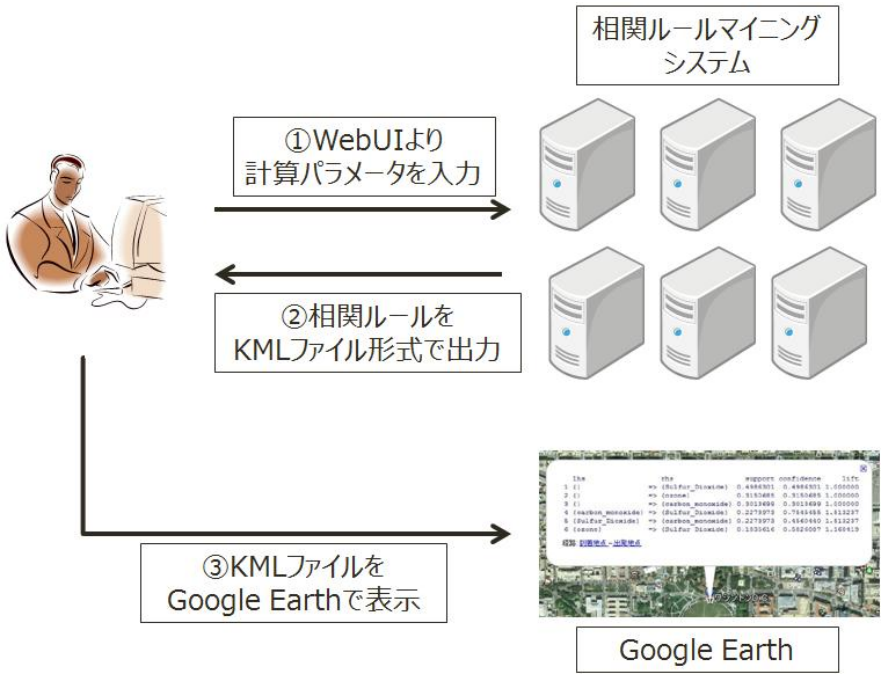


図 3-1　システム全体像

3.2　　処理の流れ

本システムは、ログデータを入力し、相関ルール分析を行い、結果をKMLファイルに出力する。また、多様なデータ形式に対応するために、プラグインを用いて入力データの中から処理に必要なデータを抽出する。相関ルール分析はHadoopと統計処理ソフトR[13]を用いて行う。

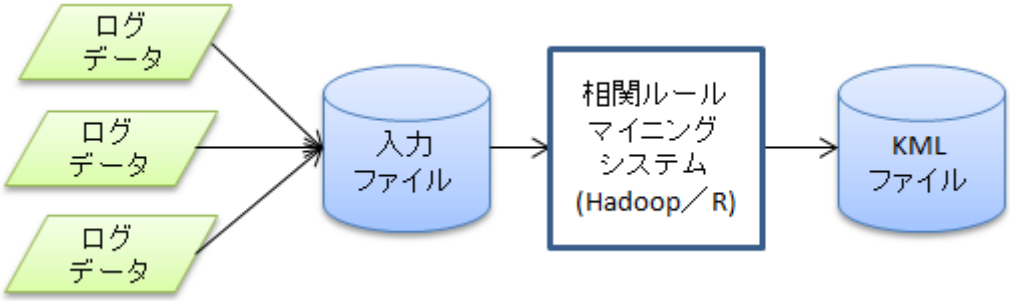


図 3-2　処理の流れ

本システムの主な機能として、データ加工機能とデータ分析機能がある。データ加工機能は異なるフォーマットのログデータを統一フォーマットに加工する。データ分析機能は、ログデータから、相関ルールを抽出し、KML ファイル形式[14]で出力する。

表 3-1　システムの主な機能

機能分類	説明
データ加工機能	入力データを中間データに加工する。多様なデータ形式に対応するために、プラグインを用いて、入力データの中から処理に必要なデータを抽出する。プラグインについては、4.2.4 で記述する。
データ分析機能	中間データから相関ルールを抽出し、KML ファイル形式で出力する。

3.3　サブシステム

システムを二つのサブシステムで構成する．一つ目はデータ加工システムである．ログデータを入力とし、中間データに変換する．中間データとは、ログデータから、分析に必要なデータを抽出したものである.二つ目はデータ分析システムである.中間データを入力とし、相関ルール分析を行い、結果を KML ファイル形式で出力する．

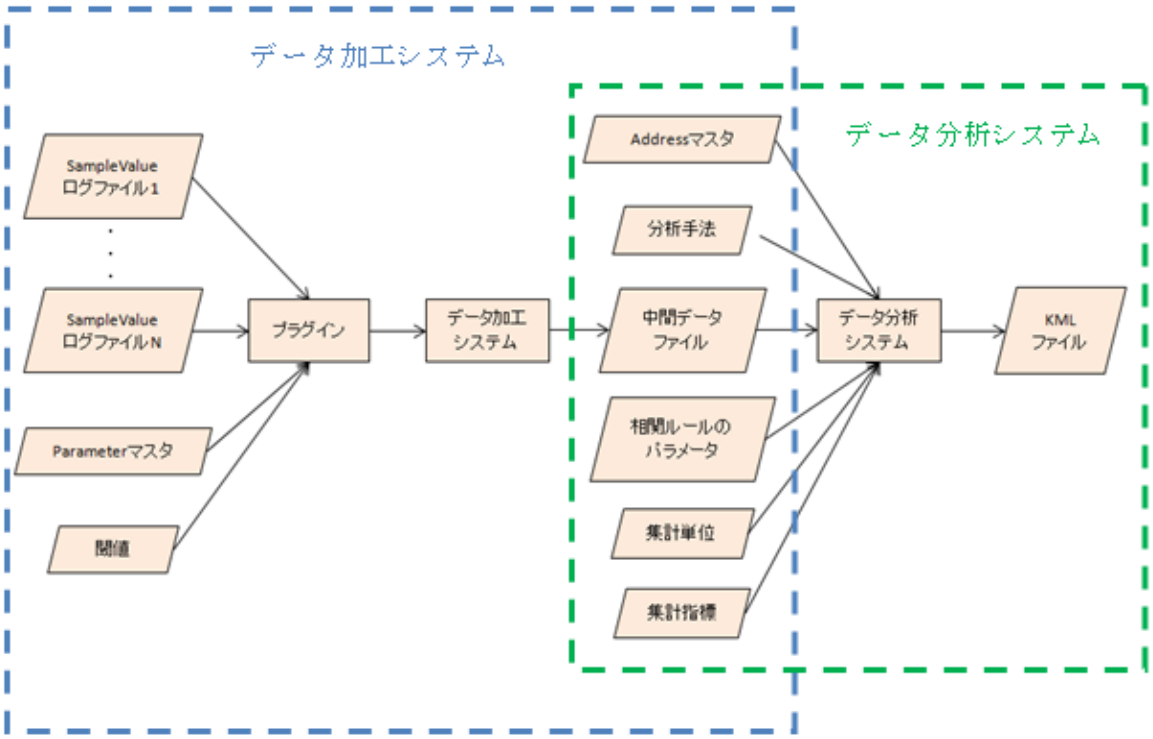


図 3-3　サブシステムの構成

3.4　　実際に扱うデータ

3.4.1　　データの調査・検討

システムの入力データについて調査・検討を行った。扱うデータの選定基準は入手しやすさとデータ量の二通りの基準を設けた。

これらの基準を元に、調査を行った結果、二つのログデータに着目した。一つ目は日本の環境省が提供している大気環境データである。このデータは大気汚染物質広域監視システム[15]から取得できる。二つ目はアメリカの環境庁が提供している大気環境データである。このデータは、Air Quality System[16]から取得できる。

表 3-2　データの比較

データ	取得方法	データ量
日本	Web	数 GB
アメリカ	Web	数十 GB

本プロジェクトでは二つのデータを比較して、データ量の大きいアメリカの環境データを用いることにする。

3.4.2　　AQS データ

米国環境保護庁(United States Environmental Protection Agency)では、WEB 上で大気汚染観測ログデータ(Air Quality System Data)を提供している。

Air Quality System (AQS)とは、EPA が所有する大気汚染観測ログデータのリポジトリである。AQS は 10,000 個以上のモニタを所有しているが、現在は 5,000 個が稼働している。大気汚染観測ログデータは、州や地域の専門機関が測定し定期的に AQS に提供している。2011 年の Nitrogen Dioxide の AQS データを図 3-4 に示す。

```
# RD|Action Code|State Code|County Code|Site ID|Parameter|POC|Sample Dura
# RC|Action Code|State Code|County Code|Site ID|Parameter|POC|Unit|Methoc
RD|I|06|019|0007|42602|1|1|008|074|20110101|00:00|8|||||||||||||||
RD|I|06|019|0007|42602|1|1|008|074|20110101|01:00|7|||||||||||||||
RD|I|06|019|0007|42602|1|1|008|074|20110101|02:00|4|||||||||||||||
RD|I|06|019|0007|42602|1|1|008|074|20110101|03:00||AX|||||||||||||
RD|I|06|019|0007|42602|1|1|008|074|20110101|04:00||BD|||||||||||||
RD|I|06|019|0007|42602|1|1|008|074|20110101|05:00|6|||||||||||||||
RD|I|06|019|0007|42602|1|1|008|074|20110101|06:00|9|||||||||||||||
RD|I|06|019|0007|42602|1|1|008|074|20110101|07:00|3|||||||||||||||
RD|I|06|019|0007|42602|1|1|008|074|20110101|08:00|3|||||||||||||||
RD|I|06|019|0007|42602|1|1|008|074|20110101|09:00|3|||||||||||||||
RD|I|06|019|0007|42602|1|1|008|074|20110101|10:00|2|||||||||||||||
RD|I|06|019|0007|42602|1|1|008|074|20110101|11:00|3|||||||||||||||
RD|I|06|019|0007|42602|1|1|008|074|20110101|12:00|5|||||||||||||||
RD|I|06|019|0007|42602|1|1|008|074|20110101|13:00|4|||||||||||||||
```

図 3-4　Air Quality System Data

本システムでは、以下の大気汚染物質の AQS データを使用する。

- Carbon Monoxide（一酸化炭素）
- Nitrogen Dioxide（二酸化窒素）
- Particulate Matter(PM10) （直径が 10μm 以下の粒子状物質）
- Particulate Matter(PM2.5) （直径が 2.5μm 以下の粒子状物質）
- Ozone（オゾン）
- Sulfur Dioxide（二酸化硫黄）

3.5 機能要件

図 3-5 と図 3-6 のように、各システムのユースケース図を作成し、機能要件を決定した(表 3-3)。

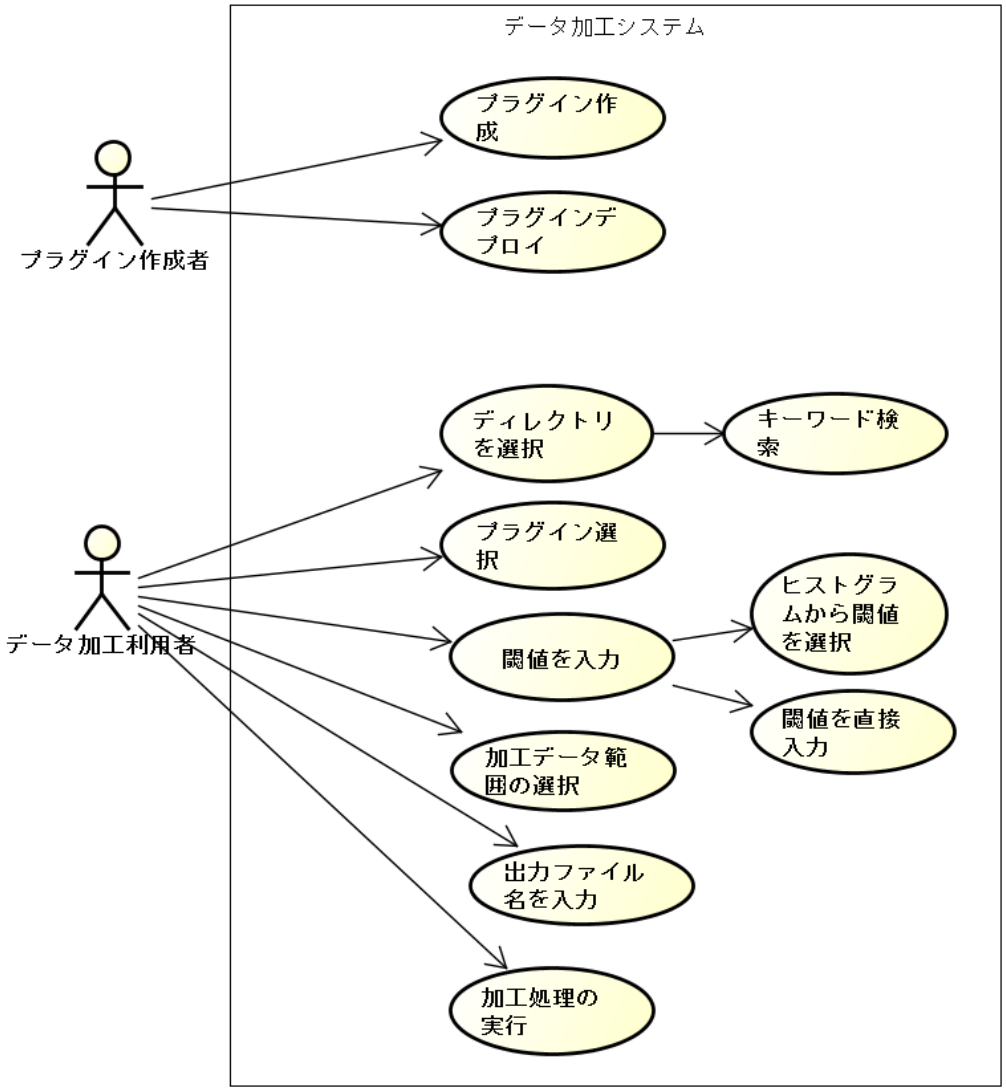


図 3-5　データ加工システムのユースケース図

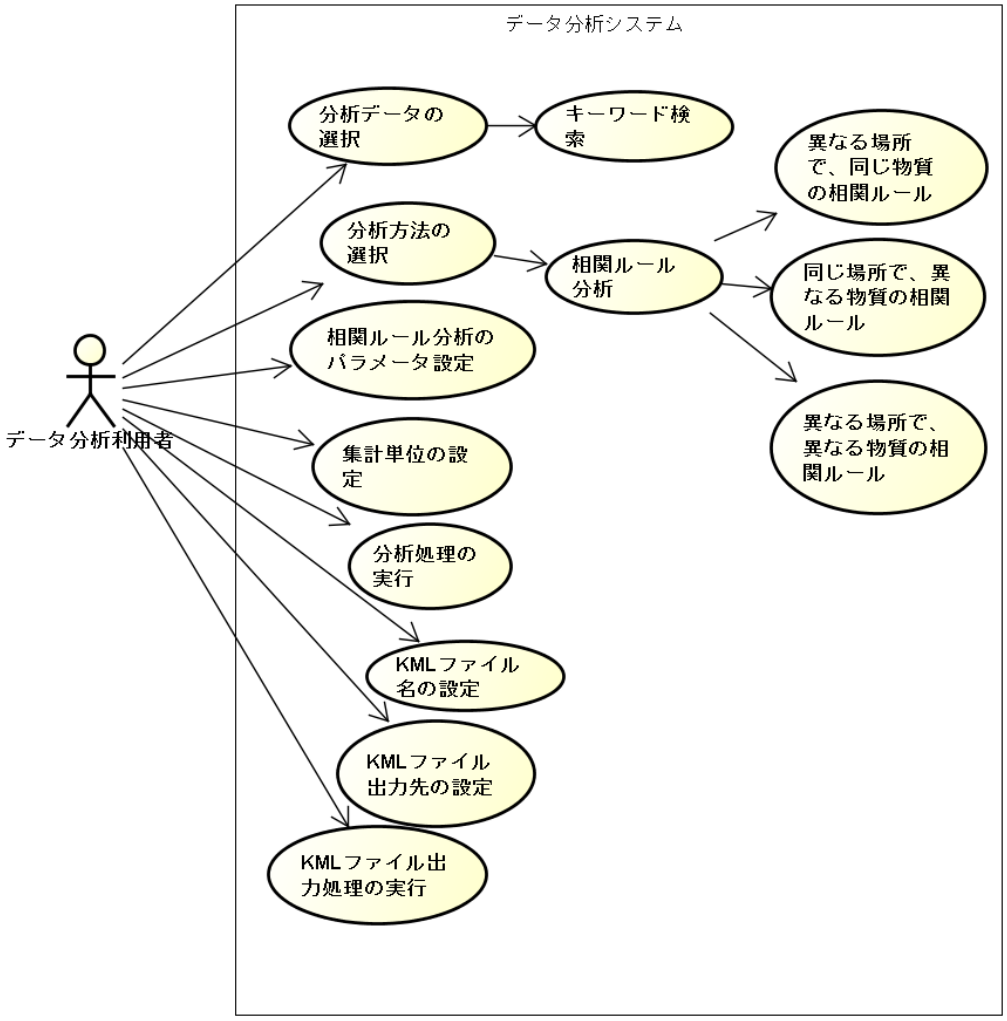


図 3-6　データ分析システムのユースケース図

表 3-3　機能要件

項番	分類	機能用件	説明
1	データ加工システム	データ入力プラグインの登録	利用者は本システムに入力プラグインを登録できる。保存先はシステム指定のフォルダ以下とする。
2		データ入力プラグインの表示	利用者は本システムに登録されているプラグインを一覧できる。プラグインはシステムが指定したフォルダ以下に入っているものとする。
3		データ入力プラグインを設定	利用者は本システムに登録されているプラグインファイルを選択できる。
4		データのフォルダを表示	利用者は本システムで加工するデータのフォルダを一覧できる。利用者は入力データフォルダのパスを指定できる。

5		データのフォルダを設定	利用者は本システムで加工するデータのフォルダを選択できる。このフォルダには全ての入力データファイルが入っている。
6		ヒストグラムを表示	利用者は本システムで加工するデータのヒストグラムを一覧できる。利用者は、ヒストグラムを参照して閾値を設定できる。
7		閾値を設定	利用者は本システムで加工するデータの閾値を、ヒストグラムに基づいて対話的に、または直接入力することで指定できる。
8		中間データファイルの出力	利用者は本システムで加工したデータの中間ファイルを取得できる。中間ファイルは、システムが指定したパスに保存する。
9	データ分析システム	中間データファイルの表示	利用者は本システムで分析するファイルを一覧できる。中間ファイルは、システムが指定したパスに入っているものとする。また、キーワードを元に該当するファイルを表示する。
10		中間データファイルの設定	利用者は本システムで分析するファイルを選択できる。
11		分析方法を設定	利用者は本システムでデータを分析する方法を選択できる。分析方法は、①異なる場所で同じ物質②同じ場所で異なる物質③異なる場所で異なる物質の３種類の相関ルールである。
12		相関ルール抽出の条件設定	利用者は本システムで相関ルールを抽出するためのパラメータ support、confidence を設定できる。
13		集計単位を設定	利用者は本システムでデータを分析する粒度を選択できる。粒度は、空間軸①州②郡③通り、または時間軸①年②月③日のいずれかである。
14		相関ルールの表示	利用者は抽出された相関ルールを画面上で一覧できる。
15		KML ファイルの保存パスを設定	利用者はKML ファイルの出力パスを設定できる。
16		KML ファイルのファイル名を設定	利用者はKML ファイルのファイル名を設定できる。
17		KML ファイルの取得	利用者は本システムで分析したデータのKML ファイルを取得できる。KML ファイルは、利用者が設定したパスにあるものとする。

3.6　想定する利用者

膨大なログデータを保持しており、そのデータを時間軸と空間軸に分割する。そして、相関ルールの分析を行い、結果を視覚化する利用者を想定している。ログデータとは、大気汚染ログデータのように、時系列で、空間情報と時間情報を含んだものを想定している。

具体的な利用者を以下に示す。

- 1 次的利用者

  研究者を想定している。膨大なログデータを保持し、そのデータを時間軸と空間軸で分割して抽出した相関ルール分析の結果を理解しうる人を想定している。ここで、ログデータとは、大気汚染ログデータのように、時系列で且つ空間情報と時間情報を含んだものを想定している。

- 2 次的利用者

  一次的利用者が作成した KML ファイルを閲覧する人を想定している。

- システム管理者

  プラグイン機能作成者を想定している。プラグイン機能作成者は、分析対象のデータを取得するためのプログラムを作成する。

3.7　　システム構成

表 3-4 にシステムに必要なソフトウェアを、図 3-7 にシステムのソフトウェア構成を示す。コンピュータの OS に Linux を採用し、その上で Hadoop と R、GoogleEarth を動作させる。また、それらの上に作成したプログラムを動作させる。

表 3-4　必要なソフトウェア

名称	製品名
OS	Linux
Java の開発環境	Java Development Kit
Java の実行環境	Java Runtime Environment
Java ソフトウェアフレームワーク	Hadoop
統計処理ソフトウェア	R
3D 地図ソフトウェア	Google Earth

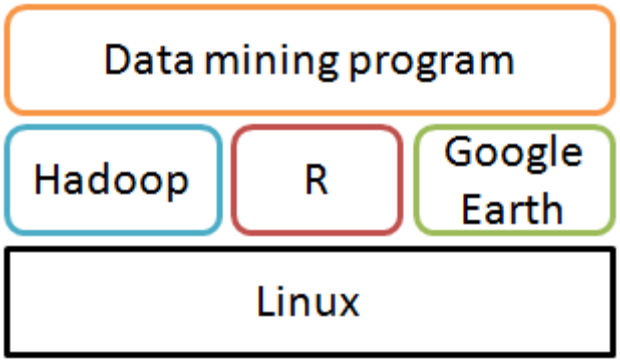


図 3-7　ソフトウェア構成

3.8　　前提条件

システムの利用者は、図 3-7 に示すソフトウェアを導入しなければならない。また、Linux は、CentOS5.6 で動作を保証する。

利用できるブラウザは、Internet Explorer、Firefox、Google Chrome とし、システムの動作を保障する。

3.9　　画面一覧

システムの画面一覧を表 3-5 に示す。

表 3-5　システムの画面一覧

画面 ID	画面名	概要
ID-1	データ加工システム操作画面	データ加工システムを操作する画面
ID-2	データ分析システム操作画面	データ分析システムを操作する画面

3.10　　画面遷移

画面遷移を図 3-8 に示す。

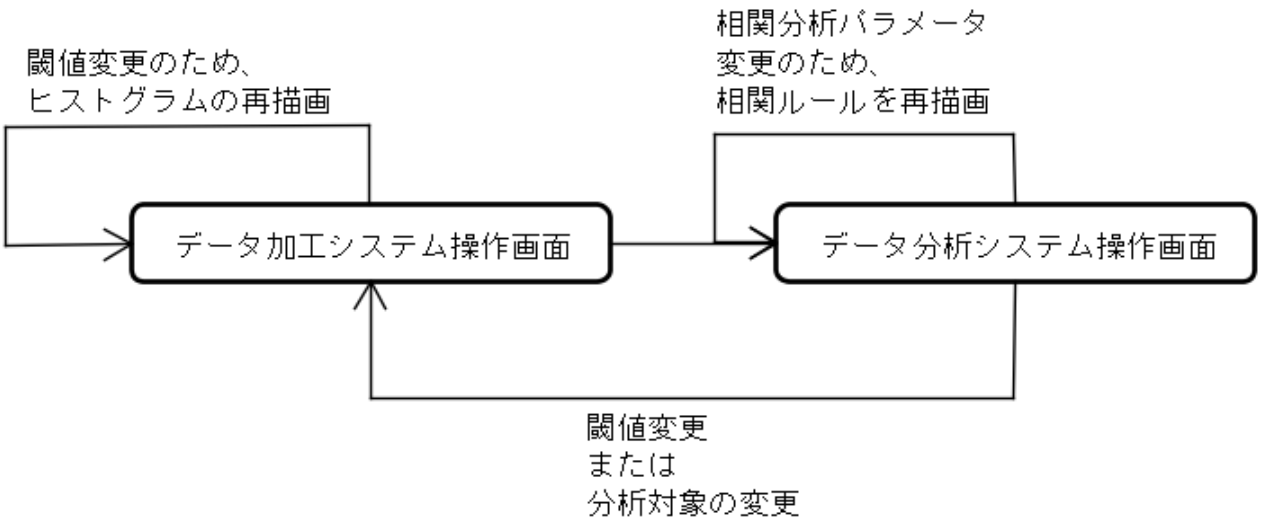


図 3-8　画面遷移図

3.11　画面構成

3.11.1　システム操作画面

図 3-9 にデータ加工システムの操作画面を示す。また、各コンポーネントの詳細を表 3-6 に示す。図 3-9 中の番号は画面項目 ID を表しており、表 3-6 中の画面項目 ID と一致する。

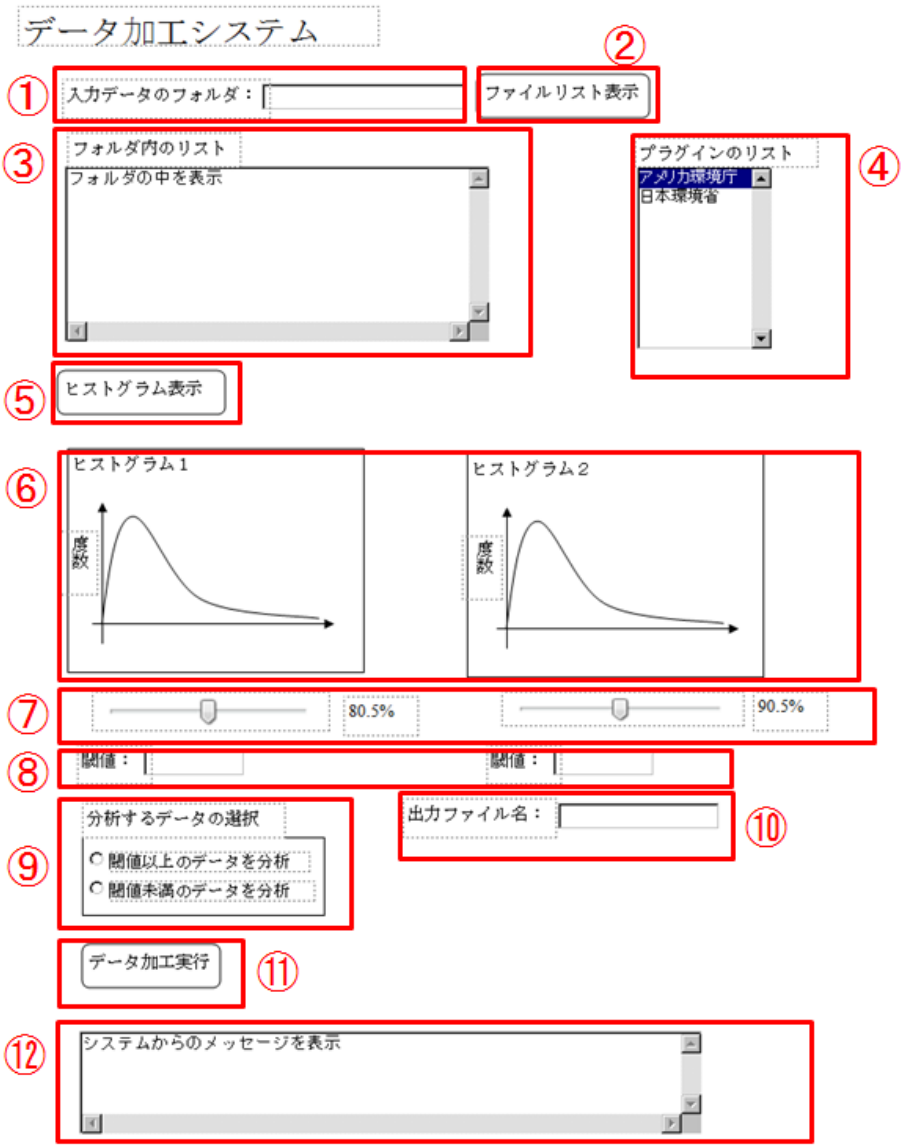


図 3-9　データ加工システム操作画面

表 3-6　各コンポーネントの詳細(データ加工システム)

画面項目ID	論理項目名	論理項目種別	データ型	入力制約	初期表示/備考
1	入力データのフォルダ	テキストボックス	String	無	空
2	ファイルリスト表示	ボタン	無	入力できない	入力フォルダ内のファイルを表示する
3	フォルダ内のリスト	リストボックス	String	入力できない	初期表示は空。入力データのフォルダ内のファイルを一覧表示する
4	プラグインのリスト	リストボックス	String	入力できない	プラグインフォルダから、プラグイン名を表示する
5	ヒストグラム表示	ボタン	無	入力できない	入力データのヒストグラムを表示する
6	ヒストグラム	図	無	入力できない	ヒストグラムの図
7	閾値のスライドバー	スライドバー	無	入力できない	スライドバーが移動すると、閾値が変化する
8	閾値	テキストボックス	String	無	スライドバーの数値を表示する
9	分析対象の選択	ラジオボタン	無	入力できない	閾値以上を分析するのか、未満を分析するのかを選択する
10	データ加工実行	ボタン	無	無	データ加工を実行する
11	出力ファイル名	テキストボックス	String	無	空
12	システムからのメッセージを表示	テキストエリア	String	入力できない	システムからのメッセージを表示する。例えば、「ヒストグラム作成中」、「データ加工実行中」など

次に、データ分析システムの操作画面を図 3-10 に示す。また、各コンポーネントの詳細を表 3-7 に示す。図 3-10 中の番号は画面項目 ID を表しており、表 3-7 中の画面項目 ID と一致する。

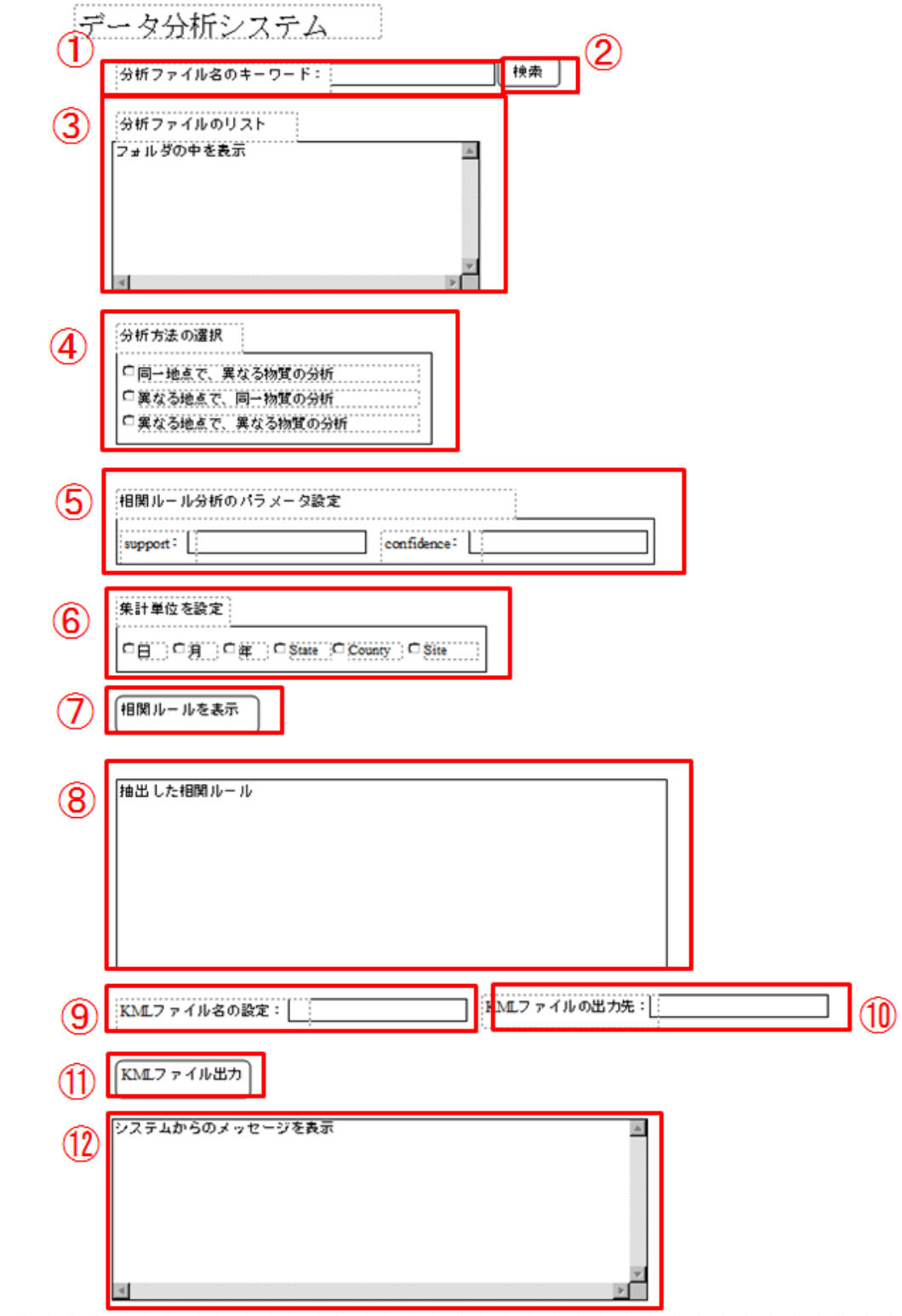


図 3-10　データ分析システムの操作画面

表 3-7　各コンポーネントの詳細(データ分析システム)

画面項目ID	論理項目名	論理項目種別	データ型	入力制約	初期表示
1	分析ファイル名のキーワード	テキストボックス	String	無	分析ファイルのキーワードを入力する
2	検索	ボタン	無	入力できない	キーワードを元に検索を行い、分析ファイルのリストに結果を表示する
3	分析ファイルのリスト	リストボックス	String	入力できない	中間データが格納されているフォルダの中から、ファイルを一覧表示する
4	分析手法の選択	ラジオボタン	無	入力できない	分析手法を選択する
5	相関ルール分析のパラメータ設定	テキストボックス	String	小数しか入力できない	設定した値以上の相関ルールを表示する
6	集計単位を設定	ラジオボタン	無	入力できない	データの粒度を設定する。設定した単位により、入力データを集計する
7	相関ルールを表示	ボタン	無	無	相関ルールを求め、表示する
8	抽出した相関ルール	テキストエリア	String	入力できない	アプリオリ関数で求めた相関ルールをそのまま表示する
9	KMLファイル名を設定	テキストボックス	String	無	KMLファイルの名前を入力する
10	KMLファイルの出力先	テキストボックス	String	無	KMLファイルの出力先を設定する
11	KMLファイル出力	ボタン	無	入力できない	KMLファイルを出力する
12	システムからのメッセージを表示	テキストエリア	String	入力できない	システムからのメッセージを表示する。例えば、「データ集計中」、「相関ルール抽出中」など

3.11.2　　システムの出力画面

本システムは、KML ファイルを出力する。なお、三つの分析方法で、Google Earth の表示内容が異なる。

【共通事項】

- 相関ルールを support の降順に表示する

相関ルールの抽出結果である support の値により、表示するアイコンを変更する。

表 3-8　アイコンと表示基準

アイコン	表示基準
	support の値が 0.75 以上
	support の値が 0.5 以上、0.75 未満
	support の値が 0.25 以上、0.5 未満
	support の値が 0.25 未満

【各分析方法固有の出力結果】

- ある場所に関する物質毎の相関ルール

ある都市において、異なる物質間での相関ルールを抽出する。Google Earth 上では、場所ごとに相関ルールを表示する。

- ある物質に関する場所毎の相関ルール

ある汚染物質について、異なる場所における相関ルールを抽出する。例えば、Nitrogen Dioxide において、ワシントンにおける汚染度が高いとラスベガスにおける汚染度も高いというルールを抽出する。Google Earth 上では条件部と結論部のアイコンを変え、その間を線で結ぶ。条件部と結論部間の線の色は、条件部と同じ色とする。また、Google Earth のフォルダ機能を用いて、汚染物質名ごとに表示の切り替えができる。

条件部のアイコン：  　　結論部のアイコン： 

- 物質と場所の相関ルール

異なる物質と場所について、相関ルールを抽出する。例えば、ワシントンの「Nirogen Dioxide」が高ければ、シアトルの「Sulfur Dioxide」も高いというルールを抽出する。Google Earth 上では、一つのルールに対して、一つ吹き出しを表示させる。

第4章　担当機能の開発

4.1　　開発機能の分担

筆者のチームでは、要件定義と設計は全員で行い、実装とテストを機能ごとに分担して行った。筆者は、データ加工システムのプラグイン機能とヒストグラム作成機能、データ加工機能を担当した。また、データ分析システムの「ある場所に関する物質ごとの相関ルール」を抽出する機能と抽出した相関ルールを場所ごとに表示する KML ファイルの作成機能も担当した。

プラグイン機能とは、多様なデータ形式に対応するために、入力データの中から処理に必要なデータを抽出するものである。プラグインに入力データの属性を保持し、入力データとプラグインのマッチングをとり、合致する入力データのみを処理対象とする。本機能は、3.5 節の項番 1 に関係する機能である。

ヒストグラム作成機能とは、利用者に表示するヒストグラムを作成する。物質ごとに度数を集計する。本機能は、3.5 節の項番 2 に関係する機能である。

データ加工システムのデータ加工機能とは、入力データを中間データのフォーマットに加工する。本機能は、3.5 節の項番 3 に関係する機能である。

「ある場所に関する物質ごとの相関ルール」を抽出する機能とは、同一の場所において、異なる物質間の相関ルールを抽出するものである。本機能は、3.5 節の項番 14 に関係する機能である。

抽出した相関ルールを場所ごとに表示する KML ファイルの作成機能とは、抽出した相関ルールを Google Earth で表示するために、相関ルールを KML ファイル形式で出力する。本機能は、3.5 節の項番 17 に関係する機能である。

表 4-1　開発担当の機能(表 3-3 より抜粋)

項番	分類	機能用件	説明
1	データ加工システム	データ入力プラグインの登録	利用者は本システムに入力プラグインを登録できる。保存先はシステム指定のフォルダ以下とする。
2		ヒストグラムを表示	利用者は本システムで加工するデータのヒストグラムを一覧できる。利用者は、ヒストグラムを参照して閾値を設定できる。
3		中間データファイルの出力	利用者は本システムで加工したデータの中間ファイルを取得できる。中間ファイルは、システムが指定したパスに保存する。
14	データ分析システム	相関ルールの表示	利用者は抽出された相関ルールを画面上で一覧できる。
17		KML ファイルの取得	利用者は本システムで分析したデータの KML ファイルを取得できる。KML ファイルは、利用者が設定したパスにあるものとする。

4.2　　データ加工システムの開発

4.2.1　　概要

本システムは複数のログデータを中間データに変換する機能を提供する。

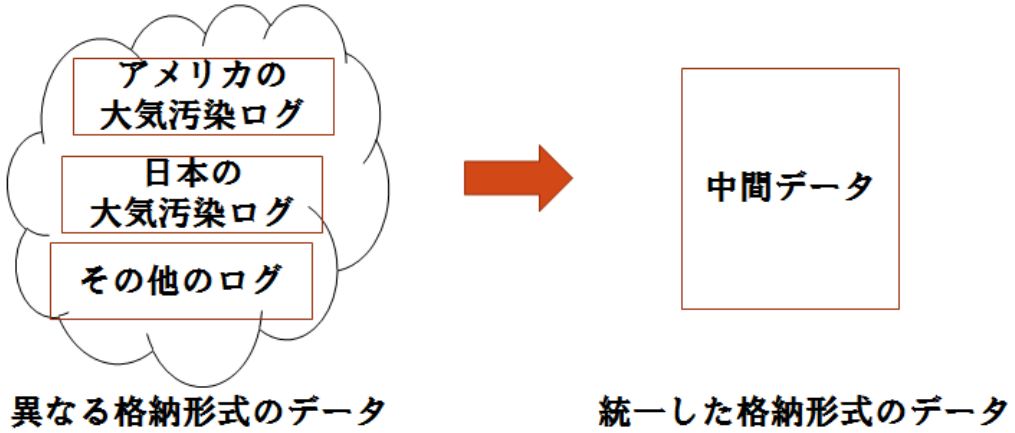


図 4-1　データ加工システム概念図

次に、システムの入出力について説明する。図 4-2 にシステムと入出力データの関係を示す。

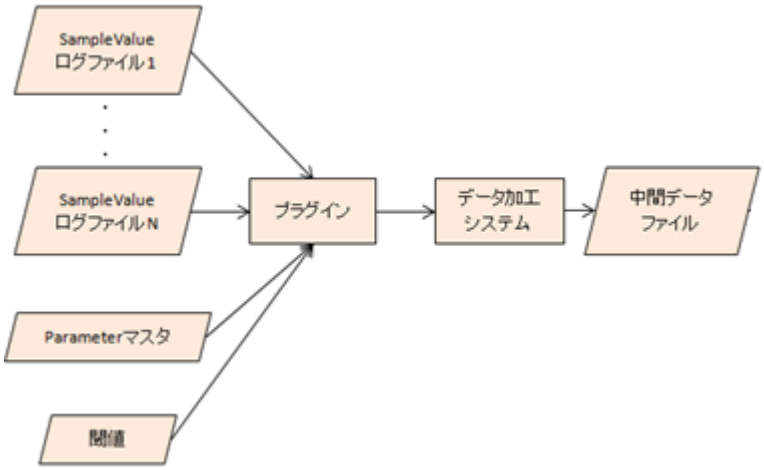


図 4-2　データ加工システム

【入力部】
システムの入力データを以下に示す。

- 大気汚染ログデータ 1~N

本システムでは、例として、大気汚染ログデータを扱う。大気汚染ログデータは csv ファイル形式でシステムに入力する。また、あらかじめ、HDFS にデータを保存しておく。

- 物質名

分析データから中間ファイルに変換するときに、大気汚染ログデータの Parameter を物質名に変えなければならない。その対応表をファイルとして入力する。

- 閾値

閾値は入力画面からシステムに入力を行う

- 有効値の範囲

有効値の範囲はマスタファイルから読み込む

【出力部】
システムは中間データファイルを出力する。出力先はシステムがあらかじめ決めたディレクトリである。

4.2.2 データ加工機能

データ加工システムには、プラグインを通して、表 4-2 のようなデータを入力する。このとき、各パラメータの意味は以下の通りである。

- State Code
  州の ID である
- County Code
  郡の ID である
- Site ID
  通りの ID である
- Parameter
  物質の ID である
- Start Time
  計測した時間である
- Sample Value
  計測した物質値である

表 4-2　入力データ例(アメリカの大気汚染ログデータ)

State Code	County Code	Site ID	Parameter	Date	Start Time	Sample Value
01	073	0028	42101	20110101	00:00	0.8
01	073	0028	42101	20110101	01:00	0.5
01	073	0028	42603	20110101	00:00	0.3
01	073	0028	42603	20110101	01:00	0.2
01	073	0028	42101	20110101	00:00	0.6

次に、出力例を示す。データ加工システムでは、計測した場所と時間により、入力データを集計する。また、Sample Value の値が閾値以上の物質を Pollutant 属性とする。なお、物質 ID(Parameter)は、このとき、物質名で置き換える。

表 4-3　出力例

Day	Start Time	State	County	Site	Pollutant
20110101	00:00	01	073	0028	Lead,Ozone,Nox
20110101	01:00	01	073	0028	Lead,Ozone

4.2.3　ヒストグラム作成機能の開発

閾値選択の補助を行うために、利用者に入力データのヒストグラムを表示する。筆者はそのヒストグラムのデータを作成する機能を実装する。実装は Hadoop を用いて行い、入力データの中から物質コードと計測値を元に、集計を行う。また、物質コードは物質名に変換する。表 4-4 に入力データ例を、表 4-5 に出力データ例を示す。

表 4-4　入力データ例(アメリカの大気汚染ログデータ)

State Code	County Code	Site ID	Parameter	Date	Start Time	Sample Value
01	073	0028	42101	20110101	00:00	0.8
01	073	0028	42101	20110101	01:00	0.5
01	073	0028	42601	20110101	00:00	0.8
01	073	0028	42603	20110101	01:00	0.5
01	073	0028	42101	20110101	00:00	0.6

表 4-5　出力例

Material Name	Sample Value	Num
42101	0.8	2
42101	0.5	1
42603	0.5	1
42101	0.6	1

4.2.4　　プラグイン機能の開発

データ加工システムでは、プラグインを通して、入力データの入出力の制御と関連データの取得を行う。プラグインは、データ加工システムに様々な入力形式のデータを対応させるためのものである。したがって、プラグインの設定のみで、対応していない入力形式のデータに対応できることが望ましい。そこで、データ加工に必要なデータ属性を決め、プラグイン内で入力データ内の該当する属性を指定する。また、指定した属性と入力データが合わない場合もあるので、データの整合を確認する仕組みも提供する。具体的に、本プロジェクトでは、以下の属性をデータ加工に必要なものとした。

- アドレス 1
  計測した場所の中で、一番目に広域な地域が格納されている属性を指定する。例えば、アメリカ大気汚染ログデータの場合は、「State Code」が該当する。
- アドレス 2
  計測した場所の中で、二番目に広域な地域が格納されている属性を指定する。例えば、アメリカ大気汚染ログデータの場合は、「County Code」が該当する。
- アドレス 3
  計測した場所の中で、三番目に広域な地域が格納されている属性を指定する。例えば、アメリカ大気汚染ログデータの場合は、「Site ID」が該当する。
- 物質コード
  物質コードが格納されている属性を指定する。例えば、アメリカ大気汚染ログデータの場合は、「Parameter」が該当する。
- 計測した日付
  計測した日付が格納されている属性を指定する。例えば、アメリカ大気汚染ログデータの場合は、「Date」が該当する。
- 計測した時間
  計測した時間が格納されている属性を指定する。例えば、アメリカ大気汚染ログデータの場合は、「Start Time」が該当する。
- 計測値
  計測した値が格納されている属性を指定する。例えば、アメリカ大気汚染ログデータの場合は、「Sample Value」が該当する。

上記の他に、プラグインの仕様を以下のように決定した。

【仕様 1】
分析対象のデータは横軸に属性、縦軸にログデータが格納されているデータのみ扱う。
また、分析対象のデータが複数ある場合は、全て同一の格納形式であるものとする。

【仕様 2】
プラグインは、以下の情報を取得できる。
- アドレスオブジェクト
- 物質属性オブジェクト
- 分析に必要な属性名とその並び順

【仕様 3】
プラグインのフォルダ構成は以下のものを想定している。
Plugin フォルダー class　プラグインの class ファイルを格納
　　－ etc　　各プラグインのマスタデータを格納

【仕様 4】
プラグインの属性名と分析対象の属性名が一致しない場合は、プラグインが不一致とする。

4.2.5 Hadoop による処理の高速化

データ加工機能とヒストグラム作成機能では、処理を高速化するために、Hadoop を用いている。Hadoop は複数の Map 処理と Reduce 処理で構成されており、それぞれを並列処理できるため、処理の高速化が図れる。本プロジェクトでは、入力データのレコードに着目しデータの分割を行い、一つのレコードを一つの Map 処理に対応させた。また、Map 処理では、一つのレコード内のある属性に着目し、属性値ごとに Reduce 処理に対応させた。着目した属性を以下に示す。

- データ加工機能
  - State Code
  - County Code
  - Site ID
  - Date
  - Start Time

- ヒストグラム作成機能
  - Parameter
  - Sample Value

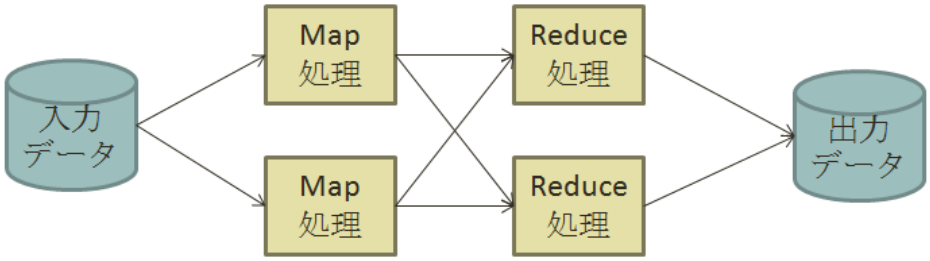


図 4-3　Hadoop による処理の高速化

4.3　ある場所に関する物質ごとの相関ルールマイニング機能の開発

4.3.1　概要

本プロジェクトでは、三つの分析を行う。筆者が開発したのは、「ある場所に関する物質ごとの相関ルール」を抽出する機能である。例えば、『ラスベガスにおいては「二酸化窒素」と「PM10」が相関関係にある』というルールを抽出できる可能性がある。

4.3.2　処理手順

処理手順を図 4-4 に示す。まず、データ加工システムの出力である中間データから、物質名を抽出し、場所ごとに物質名をまとめる。次に相関ルールを抽出する。そして、抽出した相関ルールを KML ファイル形式で出力する。

物質名を抽出時には、Hadoop を用いて処理の高速化を行っている。また、相関ルール抽出時には、統計処理ソフト R のライブラリを使用し、処理記述の簡潔化を行っている。

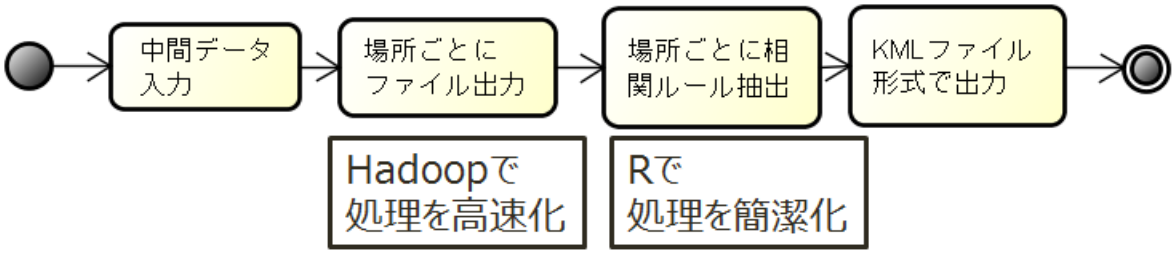


図 4-4　処理手順

4.3.3 抽出結果

図 4-5 に「ある場所に関する物質ごとの相関ルール」を抽出した結果を示す。図中の記号は以下の通りである。

表 4-6　相関ルールのパラメータ

パラメータ	意味
lhs	条件部
rhs	結論部
support	支持度であり、条件と結論を同時に満たすトランザクションが 全トランザクションに占める割合
confidence	信頼度であり、ルールの条件が発生したときに、結論が起こる割合
lift	リフト値であり、1 より大きい場合は、有効なルールといえる

図 4-5 の 1～3 行目は、頻出アイテム集合を表している。また、4 行目の結果は、carbon_monoxide と Sulfur_Dioxide に相関関係があることを表している。

```
  lhs                    rhs                    support confidence     lift
1 {}                  => {Sulfur_Dioxide}   0.4986301  0.4986301 1.000000
2 {}                  => {ozone}            0.3150685  0.3150685 1.000000
3 {}                  => {carbon_monoxide} 0.3013699  0.3013699 1.000000
4 {carbon_monoxide} => {Sulfur_Dioxide}   0.2273973  0.7545455 1.513237
5 {Sulfur_Dioxide}  => {carbon_monoxide} 0.2273973  0.4560440 1.513237
6 {ozone}           => {Sulfur_Dioxide}   0.1835616  0.5826087 1.168419
```

図 4-5　相関ルールの抽出結果

4.3.4 KML ファイルの作成

抽出した相関ルールを Google Earth 上に表示するために、図 4-6 のフォーマットの KML ファイルを作成する。

```
<xml 宣言>
<KML 開始>
        <Placemark の開始>
                <description>
                        相関ルールを HTML で記述する
                <description>
                <point の指定>
                        測定地の緯度と経度の指定
        <point の終了>
<Placemark の終了>
<KML 終了>
```

図 4-6　KML ファイルのフォーマット

第5章　データマイニングシステムの評価

5.1　　実験環境

システムの評価は 6 台のコンピュータで行い、1 台を管理コンピュータ、残りの 5 台を計算コンピュータとした。表 5-1 に、ハードウェアの性能を示す。また、コンピュータ構成を図 5-1 に示す。

表 5-1　ハードウェア性能

パソコン	CPU	Memory	台数
デスクトップパソコン	Core(TM)2 Duo CPU@2.33GHz	2GB	3
ノートパソコン	Core(TM)2 Duo CPU@1.40GHz	2GB	2
	Core(TM)2 Duo CPU@1.06GHz	2GB	1

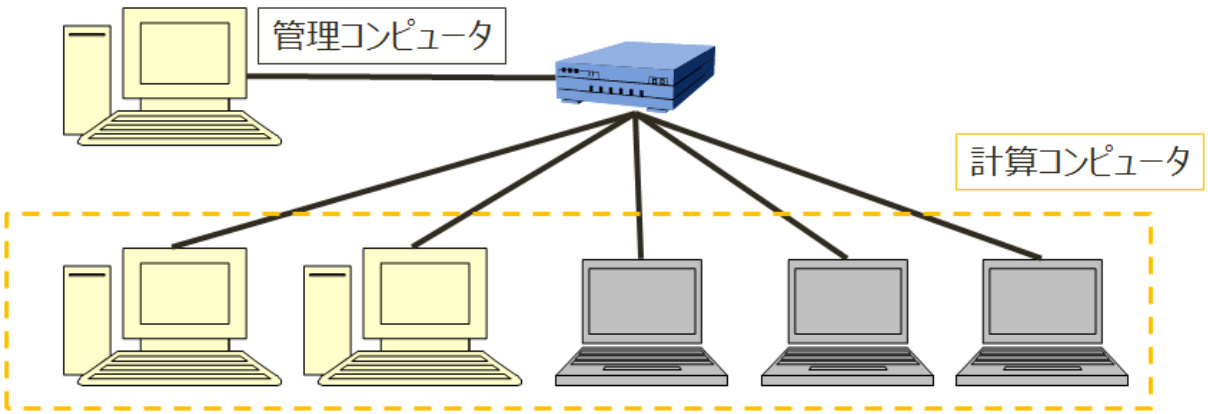


図 5-1　コンピュータ構成

次に、ソフトウェア構成を表 5-2 に示す。

表 5-2　ソフトウェア構成

名称	バージョン
OS	CentOS 5.6 Ubuntu 10.0.4
Java の開発環境	1.6.0_29
Java の実行環境	build 1.6.0_19-b11
大規模分散処理フレームワーク	Hadoop-20.0.203
統計処理ソフトウェア	2.10.0
3D 地図ソフトウェア	6.2

5.2　システムの評価項目

システムの実行速度の観点から、評価する。システムに入力するデータサイズを 5.5 G[Byte]とし、速度を比較する。

5.3　実験方法

実験手順、測定する時間を以下のように定める。

- 実験手順

① 画面からパラメータを入力し、実行ボタンを押下
② 実行後、Hadoop の Web UI より、プログラム開始時間と終了時間を取得する。

- 測定する時間
  - データ加工機能
    データ加工システムの画面上でデータ加工ボタンを押下してから、中間データを出力するまでの時間
  - データ分析機能
    データ分析システムの画面上で相関ルール表示ボタンを押下してから、分析に必要なデータが出力されるまでの時間

5.4　　入力データセット

米国環境保護庁(United States Environmental Protection Agency)が、Web 上で公開している大気汚染観測ログデータ(Air Quality System Data)を利用する。URL は http://www.epa.gov/ である。

評価には、以下の大気汚染観測ログデータを利用する。() は AQS で規定した汚染物質コードを示す。ログデータ量は合計で約 25GB である。

- Carbon Monoxide（42101）
- Sulfur Dioxide（42401）
- Nitrogen Dioxide（42602）
- Ozone（44201）
- PM-10（81102）
- PM-2.5（88101）

システムには、以下の属性を持つデータファイルを入力する。

- SampleValue.txt（ログファイル）
  Statecode, Countycode, Siteid, Prameter, Date, Starttime, Samplevalue
- Parameter.txt（マスタファイル）
  Parameter, PollutantName, minimum, maximum
- Address.txt（マスタファイル）
  StateCode, CountyCode, SiteID, StateName, CountyName, Street, Latitude, Longitude

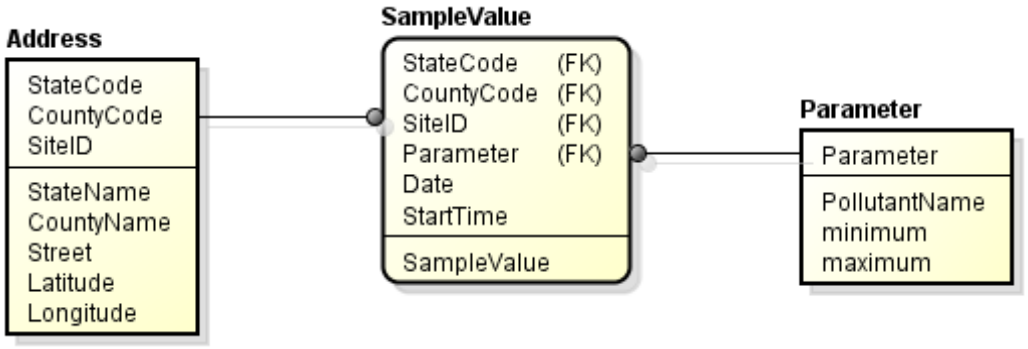


図 5-2　入力データの関係

5.5　実行速度の評価

5.4 節のデータセットを用いて、5.5G[Byte]のデータを入力し、実行速度を測定した。なお、今回は分析手法として、「ある物質に関する場所毎の相関ルール分析」を行った。測定したデータを表 5-3 に示す。

表 5-3　実行時間

(a) 計算コンピュータの台数の変化による実行速度の変化

計算コンピュータの台数	データ加工機能	データ分析機能
1 台	21 分 36 秒	5 分 3 秒
3 台	14 分 12 秒	4 分 28 秒
5 台	11 分 30 秒	3 分 52 秒

※Map 数は 2、Reduce 数は 1 に固定

(b)Map 数の変化による実行速度の変化

Map 数	データ加工機能	データ分析機能
1	11 分 52 秒	4 分 11 秒
2	11 分 30 秒	3 分 52 秒
3	10 分 27 秒	4 分 37 秒
4	10 分 13 秒	4 分 17 秒

※計算コンピュータの台数は 5 で固定

第6章　開発計画

6.1　　開発体制

プロジェクト体制を図 6-1 に示す。 委託元教員は天笠准教授であり、システムを 3 人で開発する。技術支援は日立製作所の土田正士氏と西澤格氏に行っていただく。

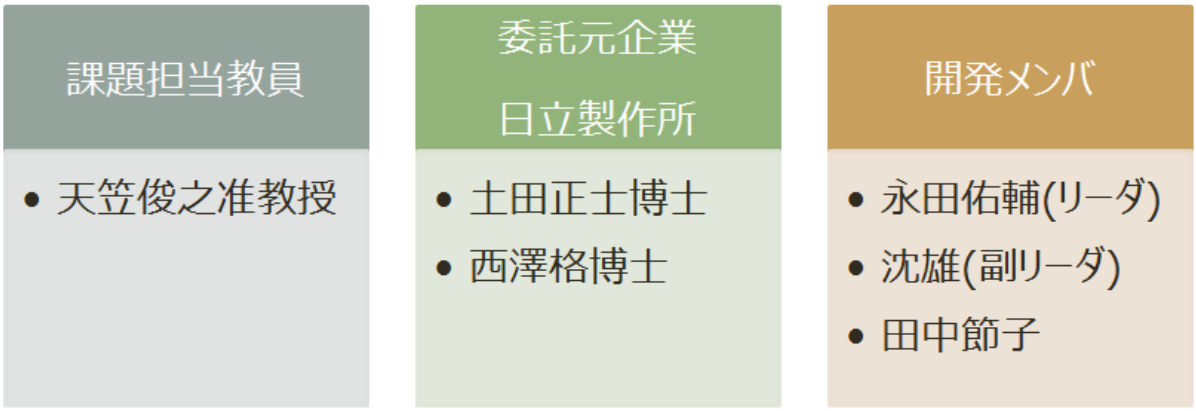


図 6-1　プロジェクト体制

次にメンバの役割を述べる。データ加工システムの設計／ 開発はメンバ全員で行う。 また、データ分析システムの設計／開発は分析方法ごとにわけ、各メンバが行う。筆者の担当は「ある場所に関する物質毎の相関ルール抽出機能の開発」である。

表 6-1　リーダと副リーダの役割

役割	担当者	作業内容
リーダ	永田	渉外、成果物のとりまとめ
副リーダ	沈	技術調査

表 6-2　作業内容と担当者

工程	作業内容	担当者
要件定義	開発構想書の作成	永田、田中、沈
設計	外部設計書の作成	永田、田中、沈
	内部設計書の作成	永田、田中、沈
データ加工システムの実装	画面イベントの開発	沈
	プラグイン機能の開発	永田
	ヒストグラムデータ作成機能の開発	永田
	中間データ作成機能の開発	永田
データ分析システムの実装	画面イベントの開発	沈
	中間データ集計機能の開発	田中
	ある場所に関する物質毎の相関ルール出力機能の開発	永田
	ある物質に関する場所毎の相関ルール出力機能の開発	田中
	物質と場所の相関ルール出力機能の開発	沈
データ加工システムのテスト	画面イベントのテスト	沈
	プラグイン機能のテスト	永田
	ヒストグラムデータ作成機能のテスト	永田
	中間データ作成機能のテスト	永田
データ分析システムのテスト	画面イベントのテスト	沈
	中間データ集計機能のテスト	田中
	ある場所に関する物質毎の相関ルール出力機能のテスト	永田
	ある物質に関する場所毎の相関ルール出力機能のテスト	田中
	物質と場所の相関ルール出力機能のテスト	沈
評価	実行速度の評価	永田、田中、沈

6.2　開発環境

本システムの開発環境として以下のものを使用する。

表 6-3　開発環境とバージョン

名称	開発環境	バージョン
OS	**CentOS**	5.6
プログラミング言語	**java**	1.6.0_29
フレームワーク	**Hadoop**	hadoop-0.20.203.0

6.3　開発スケジュール

図 6-2 に第一版の開発スケジュールを示す。本プロジェクトは、開発モデルとして、ウォータフォールモデルを採用し、要件定義、設計、実装、テスト、評価を行う。

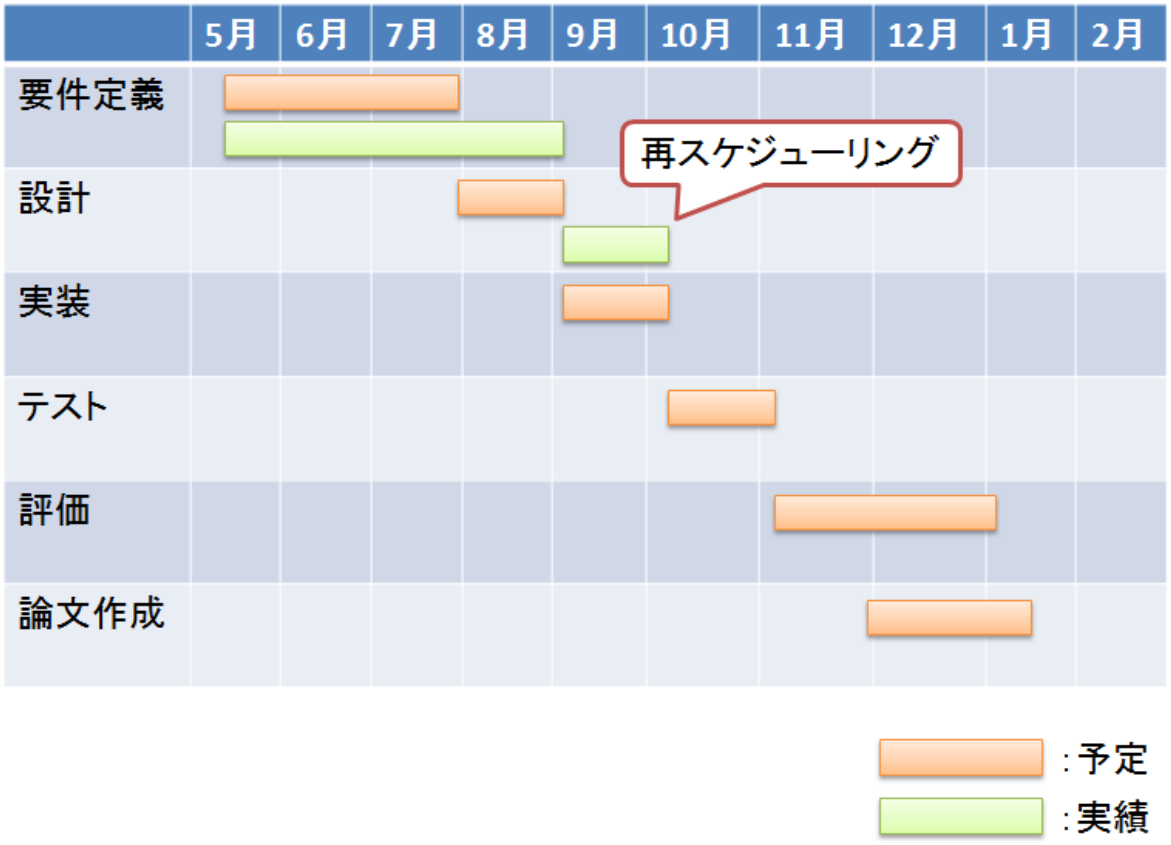


図 6-2　第一版の開発スケジュール

6.4　　開発の推移

我々は、スケジュールの遅延と、システム要件のもれより、再スケジューリングを行った。その結果を、図 6-3 に示す。

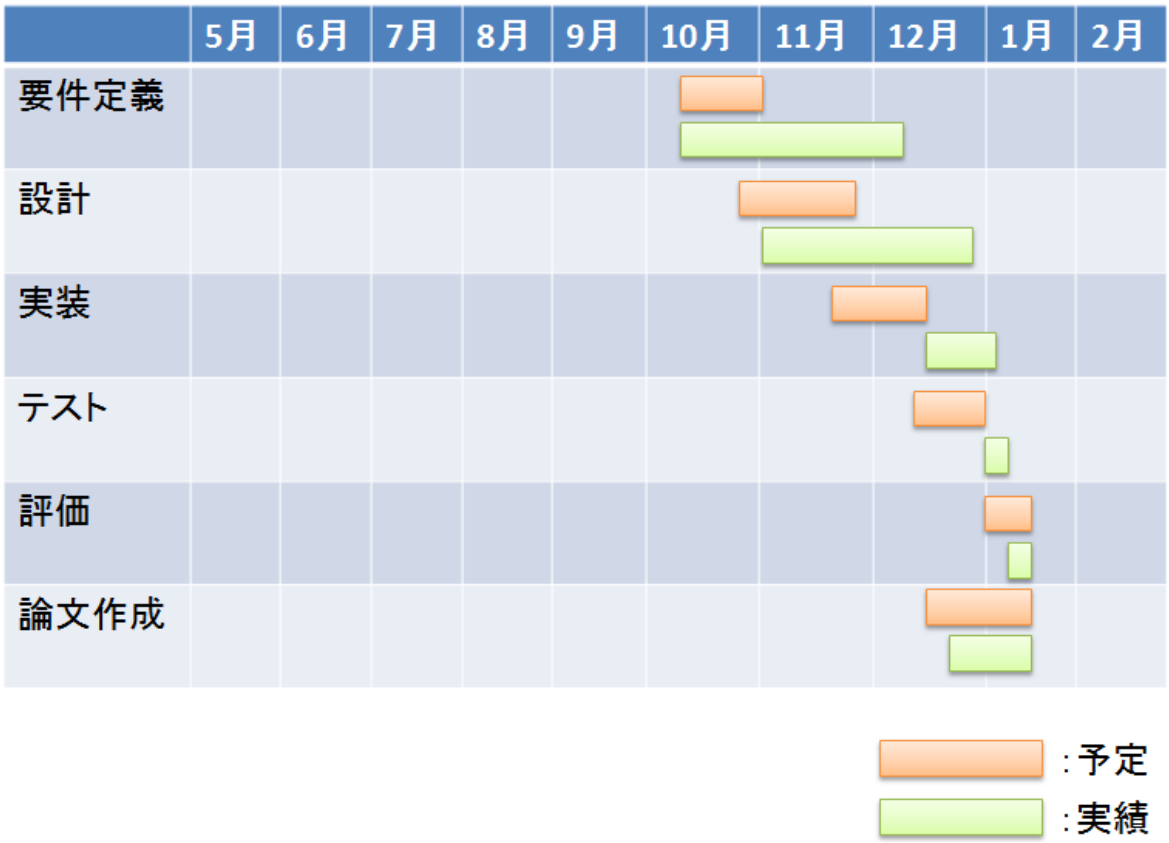


図 6-3　第二版の開発スケジュール

6.5　　各工程の成果物

本プロジェクトでは以下の文書を成果物として生成する。

- 開発構想書
- 外部設計書
- 内部設計書
- 議事録

第7章　成果物のとりまとめ

7.1　　概要

私はプロジェクトに関する成果物をまとめる作業を行っている。私がまとめた成果物は以下のものである。

- 開発構想書
- 外部設計書
- 内部設計書
- 進捗会議の資料

これらの成果物のまとめる手順を次節に示す。

7.2　　手順

成果物のとりまとめ手順を以下に示す。まず、成果物の目次構成を開発メンバで議論を行い、担当を振り分ける。そして、成果物をマージして、体裁を整える。その後、開発メンバ内で、レビューと修正を繰り返し行い、合意を得る。

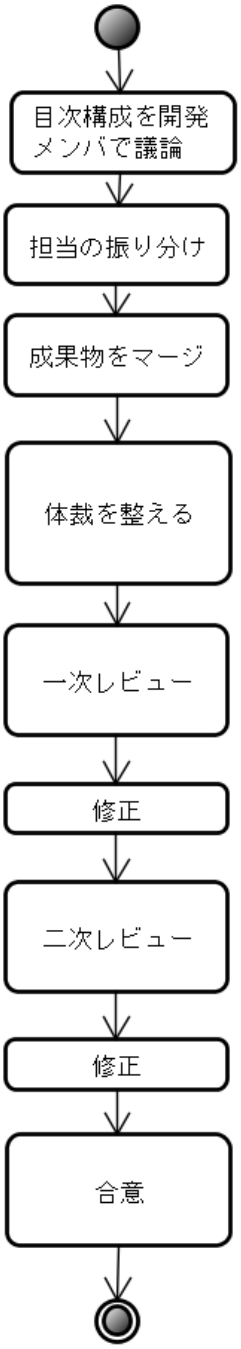


図 7-1　成果物のとりまとめ手順

7.3　　ドキュメントの作成予定期間と実績

本節では、各種ドキュメントの作成から合意までの経過を述べる。

- 開発構想書

開発構想書の目的は、プロジェクトの目的、開発するシステムの機能、その利用者を明確にすることである。開発構想書の作成から合意までの予定と実績を以下に示す。

【予定】10 月 21 日から 10 月 31 日まで
【実績】10 月 21 日から 12 月 9 日まで

- 外部設計書

外部設計書の目的は、システムの外部仕様を定義することである。外部設計書の作成から合意までの予定と実績を以下に示す。

【予定】11 月 1 日から 11 月 11 日まで
【実績】11 月 1 日から 12 月 20 日まで

- 内部設計書

内部設計書の目的は、システムの内部仕様を定義することである。内部設計書の作成から合意までの予定と実績を以下に示す。

【予定】11 月 10 日から 11 月 25 日まで
【実績】12 月 1 日から 12 月 27 日まで

7.4　　問題点と改善点

開発構想書、外部設計書、内部設計書の作成予定日数と実績は表 7-1 の通りである。

表 7-1　作成予定日数と実績

成果物	予定日数	実績
開発構想書	10 日	49 日
外部設計書	10 日	50 日
内部設計書	15 日	27 日

予定日数と実績を比較すると、大きく異なることが確認できる。これほど、遅延した原因は以下のものが考えられる。

- 事前調査(5〜8 月)に時間がかかった

- 議事録にもれがあり、次回までのアクション項目が行えなかった

- ドキュメントにあいまいな表現と定性的な表現が多く、会議を行うごとに指摘を受けていた

- レビューを行う会議の直前にドキュメントを送付していた
  レビュー時に初めてドキュメントを見ることになり、レビュー漏れが起こり、毎回、新しい箇所を指摘された

第8章　結論

研究開発プロジェクトにおいて、日立製作所が提示した「Hadoop を用いた大規模ログ分析システムの開発」というテーマの元、企画立案と仮説検証を行った。その後、設計と実装、テスト、システムの評価を行った。

企画立案では、利用できるデータを調査し、一例として、米国環境保護庁が提供している大気汚染観測ログデータを利用することを決めた。また、そのログデータを用いた相関ルール分析を行うこと決めた。仮説検証では相関ルールを抽出するパターンを三つ考え、実際に相関ルールが抽出できるか検証を行った。設計は、開発メンバ全員で行い、設計品質の向上と知識の共有化を図った。実装とテストでは、開発メンバがそれぞれ機能を分割して、行った。

筆者は、データ加工システムのプラグイン機能とヒストグラム作成機能、データ加工機能の開発を担当した。また、データ分析システムの「ある場所に関する物質ごとの相関ルール」を抽出する機能と抽出した相関ルールを場所ごとに表示する KML ファイルの作成機能の開発も担当した。成果物のとりまとめでは、成果物の作成予定日数と実績を比較した。その結果、作成予定日数に対して、実績に大きな遅れが出ていることがわかった。筆者はその理由を四つ考えた。一つ目は、事前調査(5～8 月)に時間がかかったからだと考える。二つ目は議事録にもれがあり、次回までのアクション項目が行えなかったからだと考える。三つ目はドキュメントにあいまいな表現と定性的な表現が多く、会議を行うごとに指摘を受けていたからだと考える。四つ目はレビューを行う会議の直前にドキュメントを送付していたからだと考える。

システムの評価では、Map 数と計算を行うコンピュータの台数を変化させ、実行速度を測定した。Map 数の変化させた場合、データ加工機能においては、Map 数が増えるごとに、実行時間が短縮できることが確認できた。しかし、データ分析機能においては、Map 数が増えても、実行時間はあまり変わらなかった。コンピュータの台数を変化させた場合、データ加工機能とデータ分析機能で、台数を増やすことで、実行時間の短縮が確認できた。

本プロジェクトは、会議をインターネット上で行っていたことより、作業記録を残すこととドキュメントを厳密に書くことの重要性を感じた。

謝辞

株式会社日立製作所の土田正士様、西澤格様、また、ソフトウェア事業部の皆様にはテーマの提供ならびにご指導、ご協力頂けましたことを深く感謝いたします。

本プロジェクトを進めるにあたり、委託元教員の天笠俊之准教授および指導教員の田中二郎教授には、ご指導ならびにこのような素晴らしい機会を与えて頂きました。誠に有難うございました。

本報告書執筆にあたり、ご指導頂いた高度 IT 人材育成のための実践的ソフトウェア開発専修プログラム担当の山戸昭三教授に感謝いたします。また、プロジェクトを進めるにあたり、ご指導いただいた中沢研也教授と菊池純男様に深く感謝します。

また、本プロジェクトのチームメンバである田中節子氏、沈雄氏には、プロジェクトを進めるにあたり、様々な面でのサポートを頂きました。共にプロジェクトを遂行できたことを深く感謝いたします。

最後に、様々な面で私を支えてくれた家族や多くの友人、大学生活でお世話になったすべての方々に心より感謝致します

参考文献

[1]データ・テキストマイニング、http://www.ibis.t.u-tokyo.ac.jp/yamanishi/comp.pdf
[2] DP マッチングを利用する時系列データからのデータマイニング、
http://www.ai-gakkai.or.jp/jsai/conf/2008/program/pdf/100116.pdf
[3]日経ビジネス ONLINE、http://business.nikkeibp.co.jp/article/tech/20100416/214016/
[4]The Google File System、http://www.cs.brown.edu/courses/cs295-11/2006/gfs.pdf
[5]MapReduce:Simplified Data Processing on Large Clusters、
http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/ja//archive/mapreduce-osdi04.pdf
[6] The Hadoop Distributed File Syste、
http://storageconference.org/2010/Papers/MSST/Shvachko.pdf
[7]Hadoop Streaming、http://hadoop.apache.org/common/docs/current/streaming.html
[8]Tom White、Hadoop、株式会社オライリー・ジャパン、2010 年
[9]CodeZine、http://codezine.jp/
[10]相関ルールとその周辺、
http://www.ar.sanken.osaka-u.ac.jp/~motoda/papers/or02_okada.pdf
[11]アソシエーション分析(1)、 http://mjin.doshisha.ac.jp/R/40/40.html
[12]熊谷悦生、舟尾暢男、「R」で学ぶデータマイニング Iデータ解析編、オーム社、2008 年
[13] The R Project for Statistical Computing、 http://www.r-project.org/
[14]KML、http://code.google.com/intl/ja/apis/kml/
[15]環境省大気汚染物質広域監視システム、http://soramame.taiki.go.jp/
[16] Technology Transfer Network (TTN) Air Quality System (AQS)、
http://www.epa.gov/ttn/airs/airsaqs/

付録一覧

本プロジェクトでは以下の文書を成果物として生成する。

- 開発構想書
- 外部設計書
- システム評価項目
- 議事録

日立製作所

開発構想書

～Hadoop を用いた大規模ログデータに対する相関ルールマイニングシステム～

3piece　永田佑輔　沈雄　田中節子

2011/12/01

第 1.5 版

改訂履歴

改訂番号	日付	作成者	説明
0.1	2011/10/26	永田 沈 田中	最初のバージョン
1.0	2011/10/26	永田 沈 田中	0.1 版のレビュー結果を反映した
1.1	2011/10/27	永田 沈 田中	1.0 版のレビュー結果を反映した
1.2	2011/11/07	永田	1.1 版のレビュー結果を反映した

目次

1. はじめに

1.1. 開発構想書の目的

本文書の目的は，プロジェクトの目的，開発するシステムの機能，その利用者を明確にすることである．まず，開発する目的を述べる．そして，その目的を満たすシステム要件を定義する．

なお，システムの実現方法については外部設計書と内部設計書で記述する．本文書ではシステムの要件を中心にシステムの概要を述べる．

1.2. システムの概要

1.2.1. 開発の目的

近年，社会の高度情報化と情報発信の低コスト化により，日々，大量のデータが生成されている．また，記録媒体の大容量化と通信の高速化により，膨大なデータの蓄積や流通が可能になった．そのため，企業が保有するデータ量が急激に増加している．

ビジネス環境の変化や計算機性能の向上により，膨大なデータの有効活用する試みが行われている．例えば，各種セキュリティ基準のコンプライアンスチェック，株価や交通システム利用状況などの各種分野において，大規模ログデータの利用が進んでいる．

大規模なログデータから，知識を得る方法として，データマイニングがある．データマイニングを利用することで，データの中から相関ルールのような規則性や特定のパターンを得ることができる．また，ログデータとは，毎日の気温や視聴率，商品販売数など，時間軸で連続しているデータのことで，これらデータを解析することは，未来予測や市場調査を行う上で重要である．そして，データ中のパターンの同時生起に注目した相関ルール分析の有効性が知られている．

しかし，一般的に，大規模なデータの分析処理には時間がかかる．例えば，クックパッドでは消費者の潜在的な食材へのニーズを求めるために，ユーザが入力した膨大な検索ログのデータ解析を行っており，その処理に7000時間かかると報告されている[1]．

大規模ログデータの分析処理を高速化する方法として，並列処理がある．ログデータを時間軸や空間軸を基準に分割し，処理を高速化している．また，大規模ログデータを高速に処理できる基盤として，クラウドや大規模分散処理フレームワークHadoopが注目されている．先のクックパッドの事例では，Hadoopを導入することで，処理時間を7000時間から30時間に短縮できたとも報告されている．

そこで，我々は大規模分散処理フレームワークHadoopを用いて，大規模ログに対する相関ルールマイニングシステムを開発する．また，多様なデータ形式に対応するために、プラグインを用いて入力データの入出力を制御する．

---

[1] http://business.nikkeibp.co.jp/article/tech/20100416/214016/より引用

1.2.2. 扱うログデータ

ログデータには，大気汚染の測定値のように，空間情報と時間情報を含むものがある．大気環境の分析者にとって，測定値と空間情報，時間情報を関連付けた分析結果は有益であり，環境評価／予測等に利用される．

そこで，我々は一例として本システムで扱うログデータを大気汚染ログデータとする．そして，大気汚染の測定値から，汚染物質に関する相関ルールを抽出するシステムの開発を行う．

1.2.3. システムの機能

本システムは，大気汚染ログデータを入力し，相関ルール分析を行い，結果をKMLファイルに出力する．また，プラグイン機能により，様々な分析データの格納形式にも対応する．プラグイン機能とは，様々な分析データ格納形式に，共通の関数でアクセスするためのインタフェースである．なお，プラグイン機能で扱う入力データは内部設計書で定義する．

表 1　システムが提供する機能群

機能分類	説明
データ加工機能	入力データを中間データに加工する．多様なデータ形式に対応するために、プラグインを用いて入力データの入出力を制御する。
データ分析機能	中間データから相関ルールを抽出し，KML ファイル形式で出力する．

1.3. 参考文書

[1]Tom White，Hadoop，株式会社オライリー・ジャパン，2010 年

[2]鶴保征城・駒谷昇一，『ずっと受けたかったソフトウェアエンジニアリングの授業 1』，翔泳社，2006 年

[3]鶴保征城・駒谷昇一，『ずっと受けたかったソフトウェアエンジニアリングの授業 2』，翔泳社，2006 年

2. 想定する利用者

膨大なログデータを保持しており，そのデータを時間軸と空間軸に分割する．そして，相関ルールの分析を行い，結果を視覚化する利用者を想定している．ログデータとは，大気汚染ログデータのように，時系列で，空間情報と時間情報を含んだものを想定している．

具体的な利用者は，プラグイン作成者とデータ加工／分析をする利用者の二通りを想定している． プラグイン作成者は，分析対象のデータを取得するためのプログラムを作成し，システムにデプロイする．データ加工／分析の利用者は，データから相関ルールを抽出するために，データ加工／分析を行う．

3. 開発するシステム

この節では，システムの全体像，システム構成，前提条件を述べる．

3.1. システムの全体像

このシステムは，二つのサブシステムで構成する．一つ目は分析データを入力とし，バスケットデータに変換するものであり，データ加工システムとする．また，二つ目はバスケットデータを入力とし，相関ルール分析を行い，結果をファイルに出力するデータ分析システムである．

図 1 のログデータ，プラグイン，閾値，分析手法，粒度は，システムが入力インタフェースを提供し，対話的に入力する．入力インタフェースは，Web ブラウザを利用する．なお，Web ブラウザで表示するページは，外部設計書で定義する．

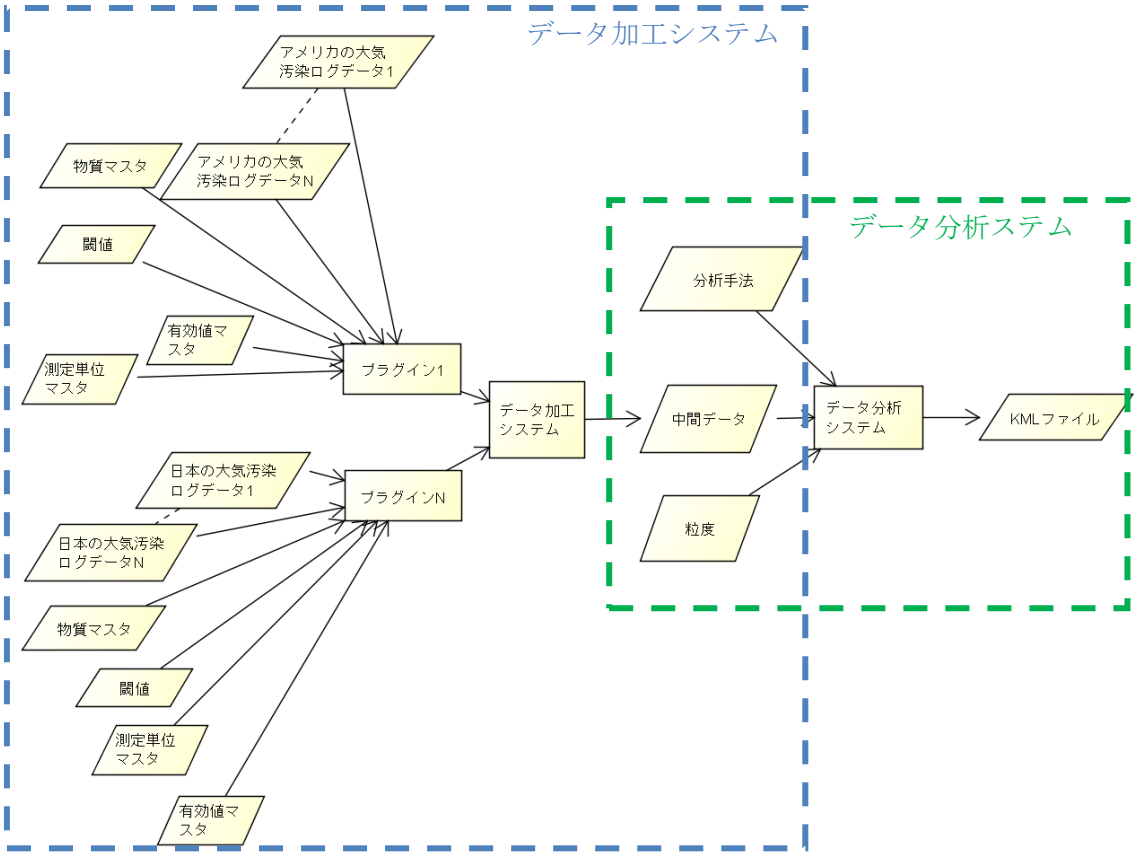


図 1　システムの全体図

3.2. システム構成

表 2 にシステムに必要なソフトウェアを，図 2 にシステムのソフトウェア構成を示す．コンピュータの OS に Linux を採用し，その上で Hadoop と R、GoogleEarth を動作させる．また，それらの上に作成したプログラムを動作させる．

表 2　必要なソフトウェア

名称	製品名
OS	Linux
Java の開発環境	Java Development Kit
Java の実行環境	Java Runtime Environment
Java ソフトウェアフレームワーク	Hadoop
統計処理ソフトウェア	R
3D 地図ソフトウェア	Google Earth

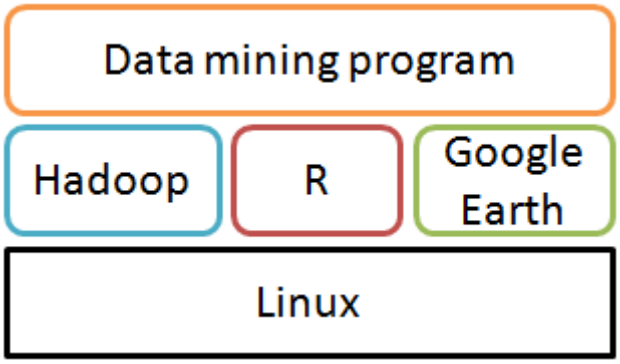


図 2　ソフトウェア構成

3.3. 前提事項

システムの利用者は，図 2 に示すソフトウェアを導入しなければならない．また，Linux は，CentOS5.6 で動作を保障する．

利用できるブラウザは，Internet Explorer，Firefox，Google Chrome とし，システムの動作を保障する．

4. 機能要件

機能要件を以下に示す.

表 3　機能要件

項番	分類	機能用件	説明
1	データ加工システム	データ入力プラグインの登録	利用者は本システムに入力プラグインを登録できる．保存先はシステム指定のフォルダ以下とする.
2		データ入力プラグインの表示	利用者は本システムに登録されているプラグインを一覧できる．プラグインはシステムが指定したフォルダ以下に入っているものとする.
3		データ入力プラグインを設定	利用者は本システムに登録されているプラグインファイルを選択できる.
4		データのフォルダを表示	利用者は本システムで加工するデータのフォルダを一覧できる．利用者は入力データフォルダのパスを指定できる.
5		データのフォルダを設定	利用者は本システムで加工するデータのフォルダを選択できる．このフォルダには全ての入力データファイルが入っている.
6		ヒストグラムを表示	利用者は本システムで加工するデータのヒストグラムを一覧できる．利用者は，ヒストグラムを参照して閾値を設定できる.
7		閾値を設定	利用者は本システムで加工するデータの閾値を，ヒストグラムに基づいて対話的に，または直接入力することで指定できる．閾値よりも汚染度の高いデータが，アソシエーション分析のバスケットに入る.
8		中間データファイルの出力	利用者は本システムで加工したデータの中間ファイルを取得できる．中間ファイルは，システムが指定したパスに保存する.

9		中間データファイルの表示	利用者は本システムで分析するファイルを一覧できる．中間ファイルは，システムが指定したパスに入っているものとする．
10		中間データファイルの設定	利用者は本システムで分析するファイルを選択できる．
11		分析方法を設定	利用者は本システムでデータを分析する方法を選択できる．分析方法は，①異なる場所で同じ物質②同じ場所で異なる物質③異なる場所で異なる物質の３種類の相関ルールである．
12	データ分析システム	相関ルール抽出の条件設定	利用者は本システムで相関ルールを抽出するためのパラメータ support，confidence を設定できる．
13		集計単位を設定	利用者は本システムでデータを分析する粒度を選択できる．粒度は，空間軸①州②郡③通り，または時間軸①年②月③日のいずれかである．
14		相関ルールの表示	利用者は抽出された相関ルールを画面上で一覧できる．
15		KML ファイルの保存パスを設定	利用者は KML ファイルの出力パスを設定できる．
16		KML ファイルのファイル名を設定	利用者は KML ファイルのファイル名を設定できる．
17		KML ファイルの取得	利用者は本システムで分析したデータの KML ファイルを取得できる．KML ファイルは，利用者が設定したパスにあるものとする．

5. ドキュメント要求

本プロジェクトでは以下の文書を成果物として生成する.

- 開発構想書(本文書)
- 外部設計書
- 内部設計書
- 試験実施報告書
- システム評価書
- 操作マニュアル
- 管理マニュアル

6. 用語集

用語集では，プロジェクトに特有の用語をすべて定義します．ユーザをはじめこの文書の読者がわからない可能性のある頭字語や略語をすべて記載します.

表 4　用語の説明

番号	用語	説明
1	相関ルール	ログデータの中から，属性間の関連を表したもの 例えば，属性 A があるならば，属性 B もあるというルールである
2	相関ルール分析	ログデータの中から，属性間の関連を分析し，相関ルールを抽出する
3	バスケットデータ	ログデータの中から，相関ルールを求めたい属性と時刻，場所を抽出したもの．閾値を元に，二値化したもの
4	support パラメータ	抽出した相関ルールの出現率である
5	confidence パラメータ	相関ルールの条件を満たしたとき，結論が起こる確率である

日立製作所　御中

外部設計書

～大規模ログデータに対する
相関ルールマイニングシステム～

筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻
3piece　永田佑輔　沈雄　田中節子

2011/12/27

第 0.1 版

目次

1. はじめに

1.1. システム名称

本システムの名称は

Hadoop を用いた大規模ログデータに対する相関ルールマイニングシステム

である．

1.2. システム概要

本システムは，大気汚染ログデータを入力し，相関ルール分析を行い，結果を KML ファイルに出力する．また，プラグイン機能により，様々な分析データの格納形式にも対応する．プラグイン機能とは，様々な分析データ格納形式に，共通の関数でアクセスするためのインタフェースである．なお，プラグイン機能で扱う入力データは内部設計書で定義する．

表 1　システムが提供する機能群

機能分類	説明
データ加工機能	入力データを中間データに加工する．プラグイン機能により，様々な入力データの格納形式にも対応する．
データ分析機能	中間データから相関ルールを抽出し，KML ファイル形式で出力する．

2. 利用者の操作とシステムの処理

利用者の操作とシステムの処理をシステムごとに記述する.

2.1. 図の凡例

凡例を示す.

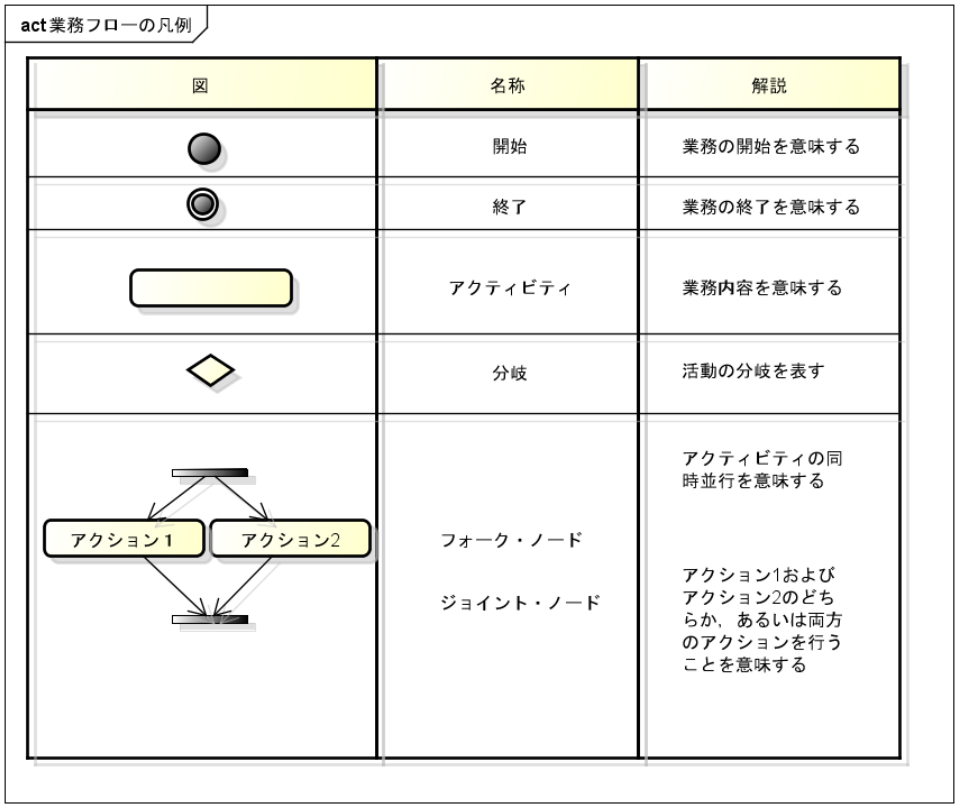


図 1 業務フローの凡例

2.2. データ加工システム

画面上での利用者の操作と対応するシステムの処理を図 2 に示す.

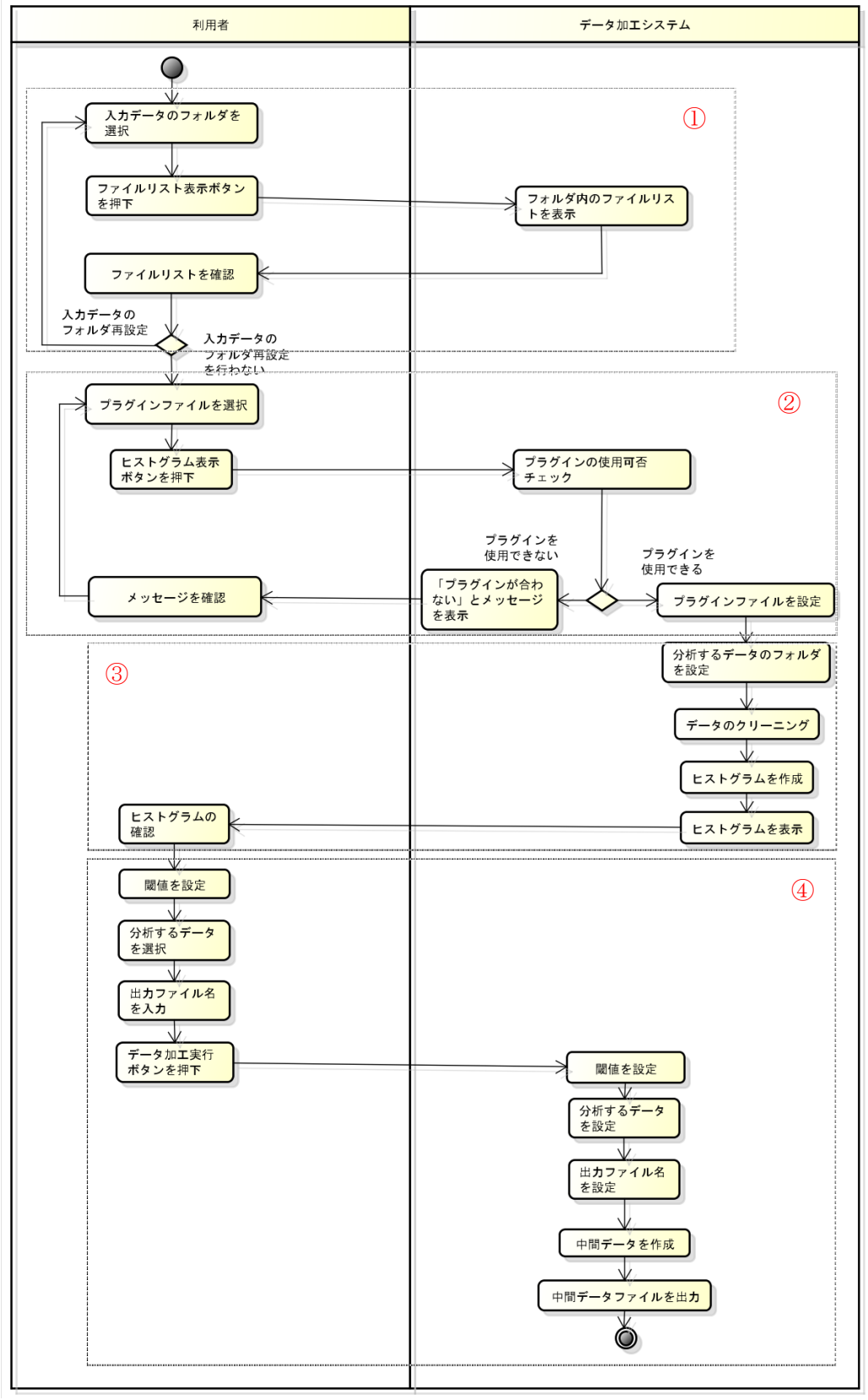


図 2　利用者の操作と対応するシステムの処理(データ加工システム)

① ファイルリスト表示の処理

利用者が入力データのフォルダを選択する。システムが入力データのフォルダ名を受け取り、HDFS 内にあるフォルダ内を表示する。利用者はファイル名を確認する。

② プラグイン選択の処理

利用者がプラグインを選択し、ヒストグラム表示ボタンを押下する。システムは、プラグインのチェックを行う。チェックの結果、プラグインが使用できなければ、メッセージを表示する。

③ ヒストグラム表示の処理

データを取得し、クリーニングを行う。そして、ヒストグラムの作成を行う。

④ 中間データファイル作成の処理

利用者は閾値と分析するデータ、出力するファイル名を入力する。システムは入力値を元に、データ加工を行い、HDFS 内に指定したファイル名で出力する。

2.3. データ分析システム

画面上での利用者の操作と対応するシステムの処理を図 3 に示す.

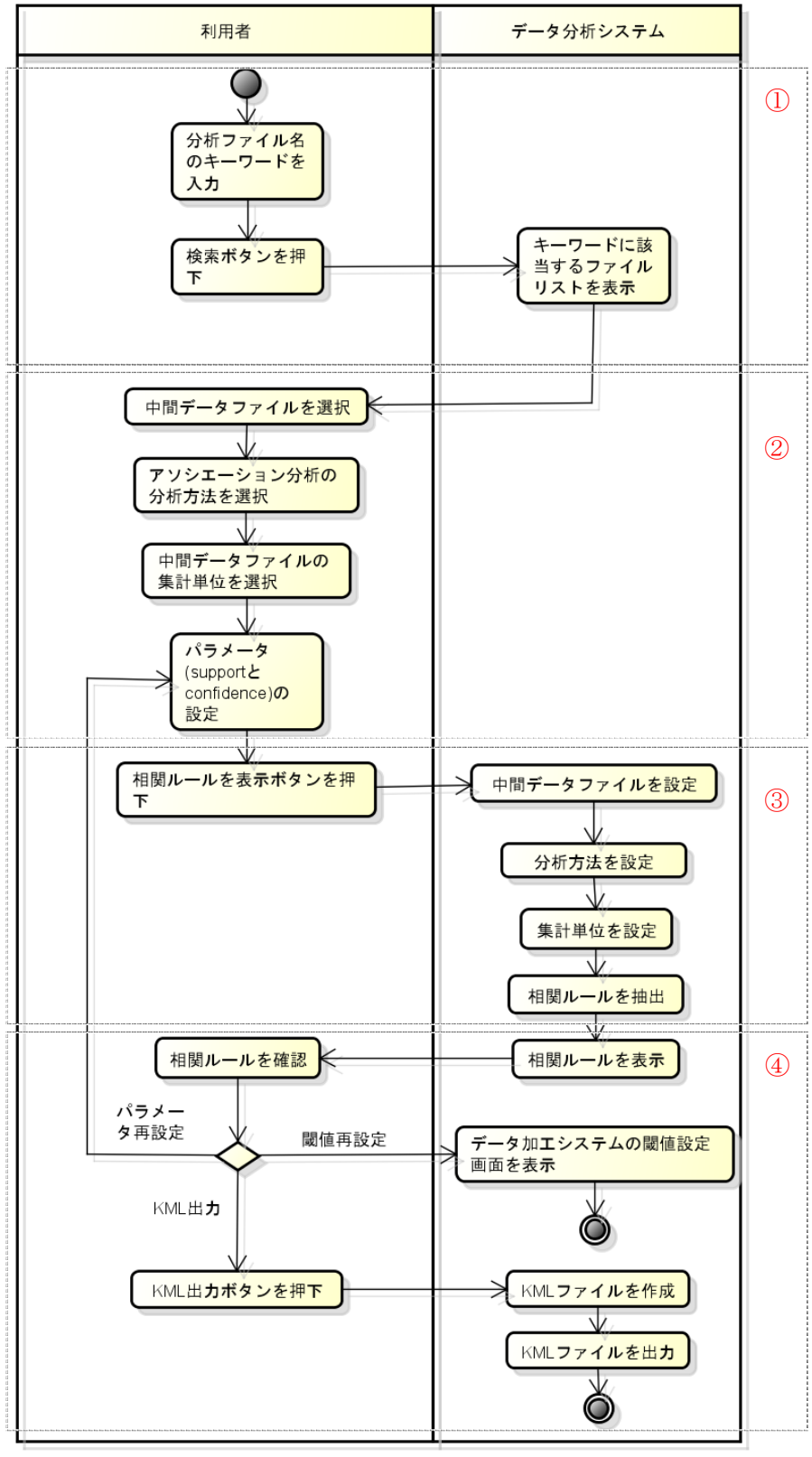


図 3　利用者の操作と対応するシステムの処理(データ分析システム)

① 中間データファイルのリスト表示処理

利用者は中間データファイルを検索するためのキーワードを入力し、検索ボタンを押下する。システムはキーワードを受け取り、ファイル名にキーワードを含む中間データファイルのリストを表示する。利用者は中間データファイルのリストを確認する。

② 中間データファイル集計のためのパラメータ設定処理

利用者は、中間データファイル、アソシエーション分析の分析方法、中間データファイルの集計方法、相関ルール抽出のパラメータを設定し、相関ルール表示ボタンを押下する。

③ 相関ルール抽出の処理

システムは、各パラメータを受け取り、相関ルールを抽出する。そして、抽出された相関ルールを表示する。

④ KML ファイル作成の処理

利用者は、相関ルールを確認し、ローカルディスク上のパス、ファイル名を入力し、KML 出力ボタンを押下する。システムは、パス、ファイル名を受け取り、KML ファイルを作成し保存する。利用者は、中間データファイル集計のためのパラメータを再設定し、相関ルールの抽出を行うこともできる。

3. システム構成

3.1. ハードウェア構成

表 2 にシステムに必要なソフトウェアを示す．コンピュータの OS に Linux を採用し，その上で Java の実行環境と R を動作させる．また，Java の実行環境の上に Hadoop を，Hadoop の上に作成したプログラムを動作させる．

表 2　必要なソフトウェア

名称	製品名
OS	Linux
Javaの開発環境	Java Development Kit
Javaの実行環境	Java Runtime Environment
Javaソフトウェアフレームワーク	Hadoop
統計処理ソフトウェア	R
3D地図ソフトウェア	Google Earth

3.2. ソフトウェア構成

図 4 にシステムに本システムのソフトウェア構成を示す．オレンジ枠の部分は今回開発するプログラムである．

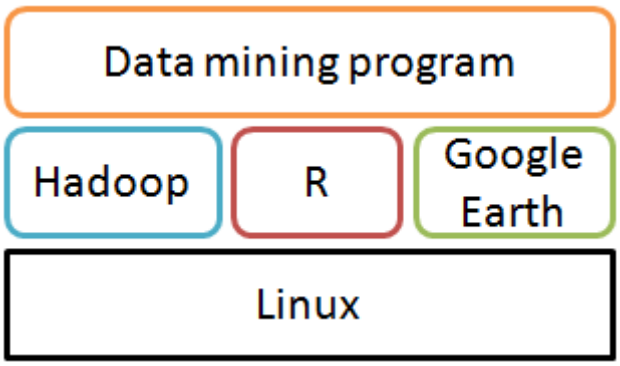


図 4　ソフトウェア構成

4. ユーザインターフェース

4.1. 画面一覧

4.1.1. システム画面

システムの画面一覧を以下の表に示す.

表 3　システムの画面一覧

画面 ID	画面名	概要
ID-1	データ加工システム操作画面	データ加工システムを操作する画面
ID-2	データ分析システム操作画面	データ分析システムを操作する画面

4.2. 画面遷移

画面遷移に関しては，別資料の画面遷移に詳細を示す.

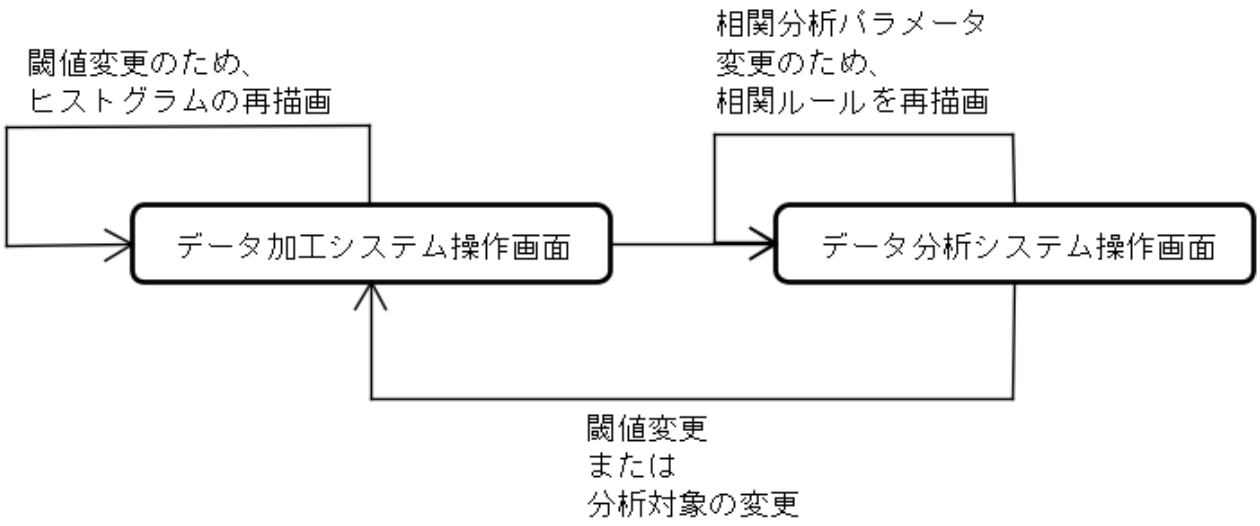
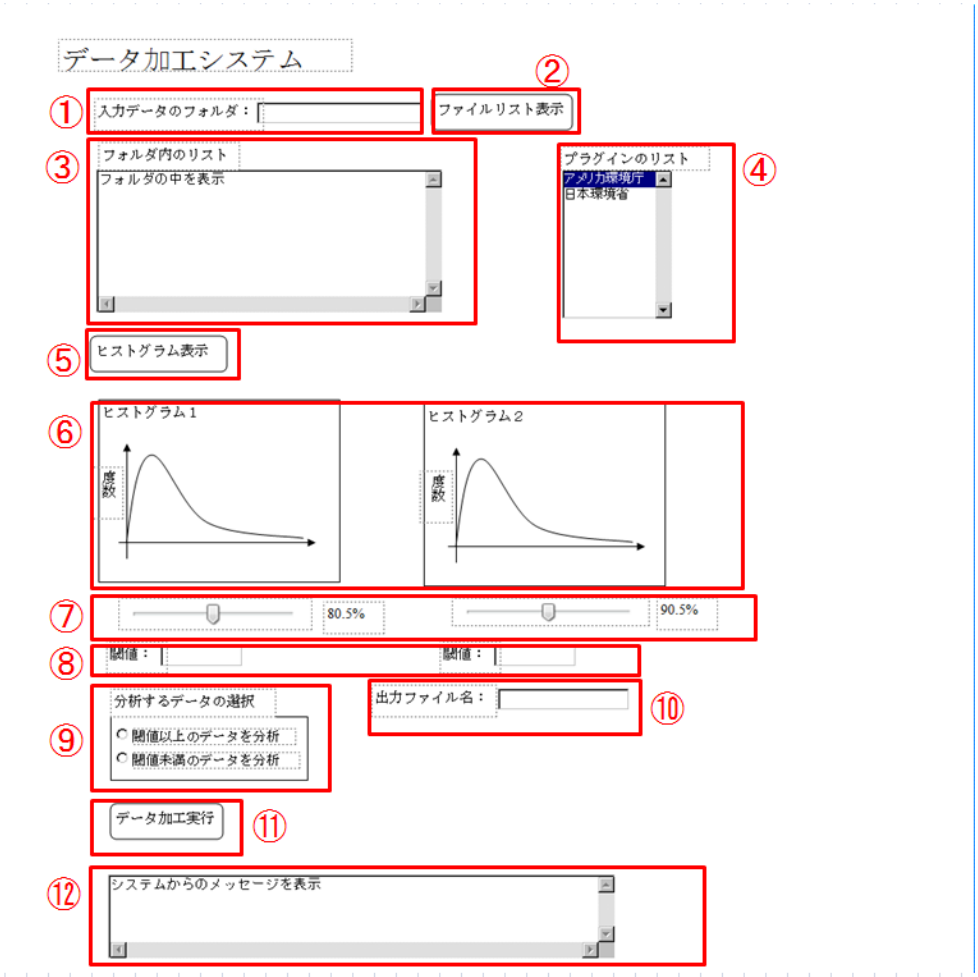


図 5 画面遷移図

4.3. 画面構成

4.3.1. システム操作画面

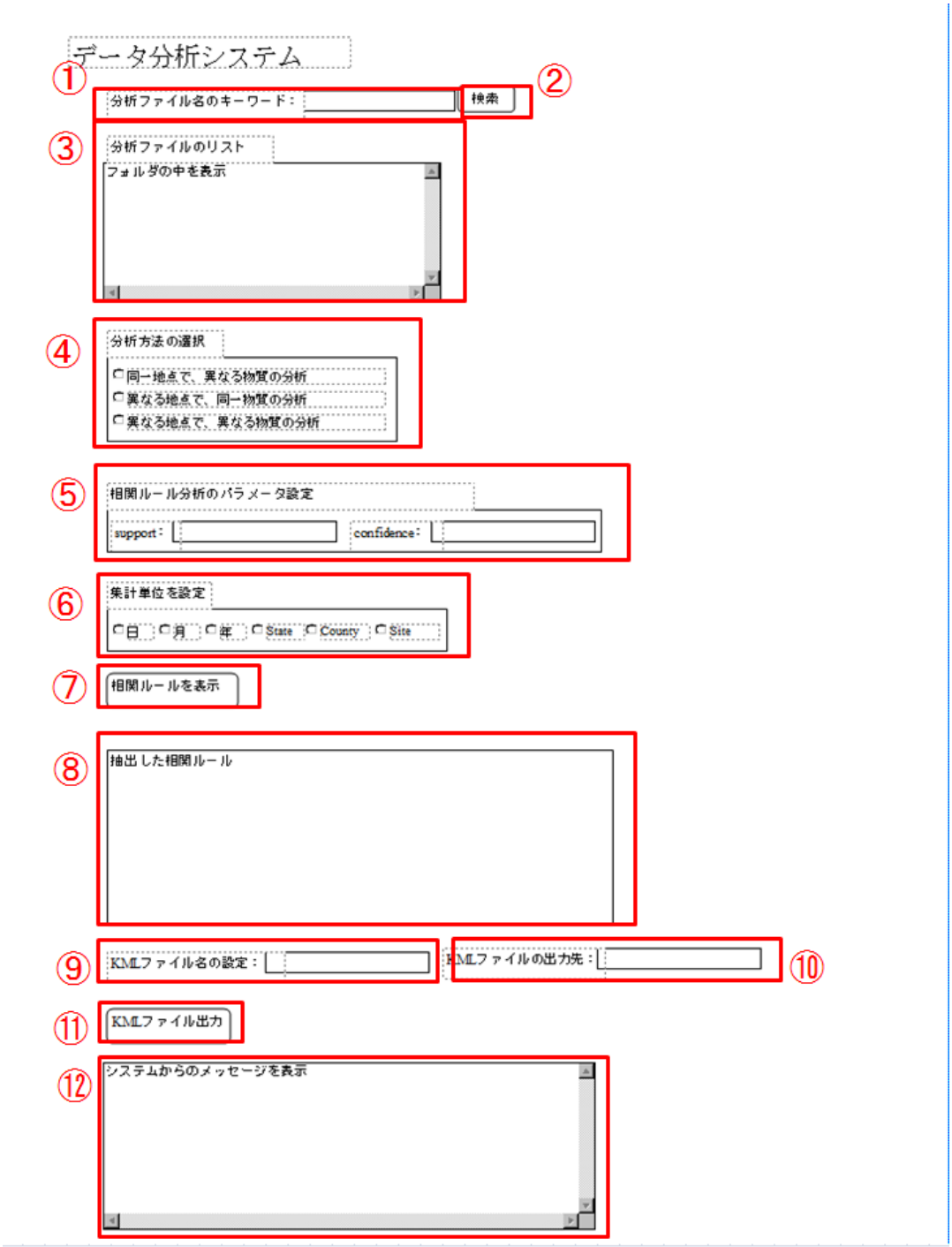
図 6 にデータ加工システムの操作画面を示す。



画面項目 ID	論理項目名	論理項目種別	データ型	入力制約	初期表示/備考
1	入力データのフォルダ	テキストボックス	String	無	空
2	ファイルリスト表示	ボタン	無	入力できない	入力フォルダ内のファイルを表示する
3	フォルダ内のリスト	リストボックス	String	入力できない	初期表示は空。入力データのフォルダ内のファイルを一覧表示する
4	プラグインのリスト	リストボックス	String	入力できない	プラグインフォルダから、プラグイン名を表示する
5	ヒストグラム表示	ボタン	無	入力できない	入力データのヒストグラムを表示する
6	ヒストグラム	図	無	入力できない	ヒストグラムの図
7	閾値のスライドバー	スライドバー	無	入力できない	スライドバーが移動すると、閾値が変化する
8	閾値	テキストボックス	String	無	スライドバーの数値を表示する
9	分析対象の選択	ラジオボタン	無	入力できない	閾値以上を分析するのか、未満を分析するのかを選択する
10	データ加工実行	ボタン	無	無	データ加工を実行する
11	出力ファイル名	テキストボックス	String	無	空
12	システムからのメッセージを表示	テキストエリア	String	入力できない	システムからのメッセージを表示する。例えば、「ヒストグラム作成中」、「データ加工実行中」など

図 6　システム操作画面と各コンポーネント（データ加工システム）

図 7 にデータ分析システムの操作画面を示す。



画面項目ID	論理項目名	論理項目種別	データ型	入力制約	初期表示
1	分析ファイル名のキーワード	テキストボックス	String	無	分析ファイルのキーワードを入力する
2	検索	ボタン	無	入力できない	キーワードを元に検索を行い、分析ファイルのリストに結果を表示する
3	分析ファイルのリスト	リストボックス	String	入力できない	中間データが格納されているフォルダの中から、ファイルを一覧表示する
4	分析手法の選択	ラジオボタン	無	入力できない	分析手法を選択する
5	相関ルール分析のパラメータ設定	テキストボックス	String	小数しか入力できない	設定した値以上の相関ルールを表示する
6	集計単位を設定	ラジオボタン	無	入力できない	データの粒度を設定する。設定した単位により、入力データを集計する
7	相関ルールを表示	ボタン	無	無	相関ルールを求め、表示する
8	抽出した相関ルール	テキストエリア	String	入力できない	アプリオリ関数で求めた相関ルールをそのまま表示する
9	KMLファイル名を設定	テキストボックス	String	無	KMLファイルの名前を入力する
10	KMLファイルの出力先	テキストボックス	String	無	KMLファイルの出力先を設定する
11	KMLファイル出力	ボタン	無	入力できない	KMLファイルを出力する
12	システムからのメッセージを表示	テキストエリア	String	入力できない	システムからのメッセージを表示する。例えば、「データ集計中」、「相関ルール抽出中」など

図７　システム操作画面と各コンポーネント(データ分析システム)

4.4. イベント一覧

以下にイベント一覧を示す.

表 4　データ加工システム操作画面

番号	イベント名	遷移先	処理内容
1	ファイルリストの表示	同画面	指定したフォルダのファイルリストを表示する
2	ヒストグラムの表示	同画面	入力データのヒストグラムを表示する
3	データ加工実行	データ分析システム操作画面	入力データから、中間データを作成する

表 5　データ分析システム操作画面

番号	イベント名	遷移先	処理内容
1	分析データの検索	同画面	中間データを格納しているフォルダから、キーワードに関連するファイルを検索する
2	相関ルールを表示	同画面	相関ルールを抽出し、表示する
3	KML ファイル出力	同画面	指定したフォルダに KML ファイルを出力する

5. 機能説明

本節では，ユースケースを用いてシステムの機能を詳細に定義する．

5.1. ユースケース一覧

以下の表にユースケース一覧を示す．

表 6　ユースケース一覧(データ加工システム)

No	アクター	ユースケース名	ユースケース概要
1	プラグイン作成者	プラグイン作成	入力データを読み取るためのプラグインを作成する
2	プラグイン作成者	プラグインのデプロイ	作成したプラグインを指定のディレクトリに格納する
3	データ加工利用者	プラグインの選択	該当するプラグインを選択する
4	データ加工利用者	ディレクトリの入力	中間データのディレクトリを入力する
5	データ加工利用者	閾値の入力	二値化の基準値を入力する
6	データ加工利用者	加工データ範囲の選択	閾値より大きい値を加工するか, 小さい値を加工するか選択する
7	データ加工利用者	出力ファイル名を入力	出力ファイル名を入力する
8	データ加工利用者	加工処理の実行	指定したディレクトリ内のファイルを中間データファイルに変換する

表 7　ユースケース一覧(データ分析システム)

No	アクター	ユースケース名	ユースケース概要
1	データ分析利用者	分析データの選択	分析を行う中間データファイルを選択する
2	データ分析利用者	分析方法の選択	三つの分析方法から選択する
3	データ分析利用者	相関ルール分析のパラメータ設定	相関ルール分析のパラメータ設定(confidence と support)を行う
4	データ分析利用者	集計単位の設定	分析する時の粒度を設定する. 本システムでは，時間(日／月／年)と空間(/county／state)単位で，データをまとめ，分析を行うことができる
5	データ分析利用者	分析処理の実行	指定した中間データファイルの分析処理を実行する
6	データ分析利用者	KML ファイル名の設定	KML ファイル名の設定を行う
7	データ分析利用者	KML ファイル出力先の設定	KML ファイル出力先の設定を行う
8	データ分析利用者	KML ファイルの出力処理の実行	KML ファイルの出力処理を実行する

5.2. ユースケース図

システムの利用シーン別に，ユースケース図を示す.

5.2.1. ユースケース図の凡例

ユースケース図の凡例を示す.

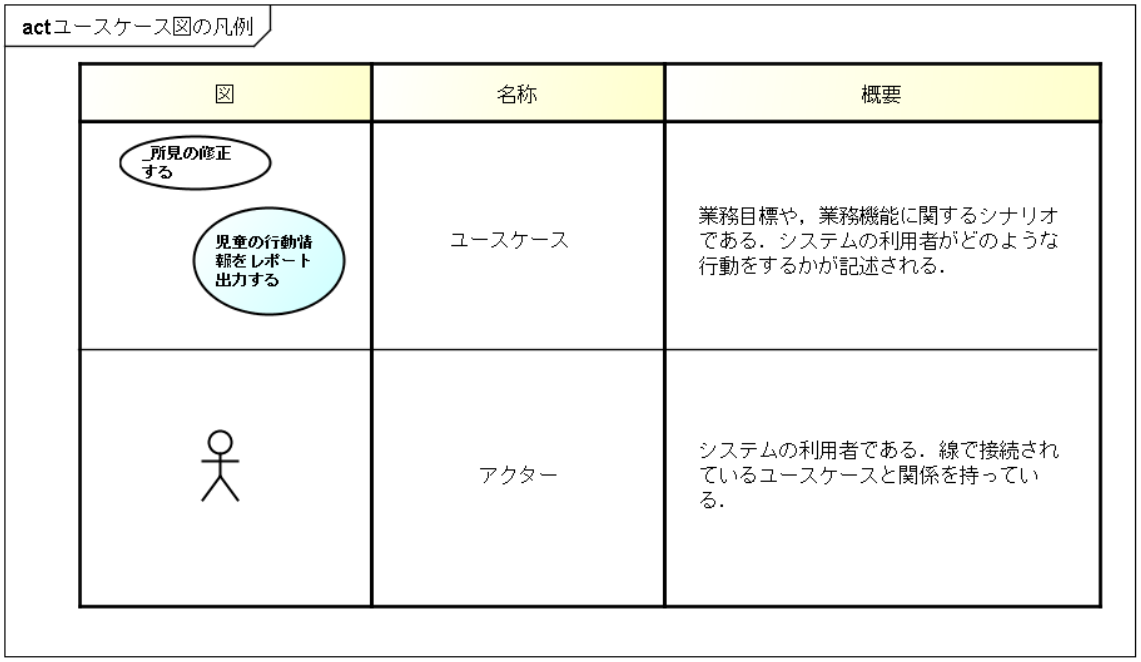


図 8 ユースケース図の凡例

5.2.2. データ加工システム

データ加工システムのユースケースを図 9 に示す.

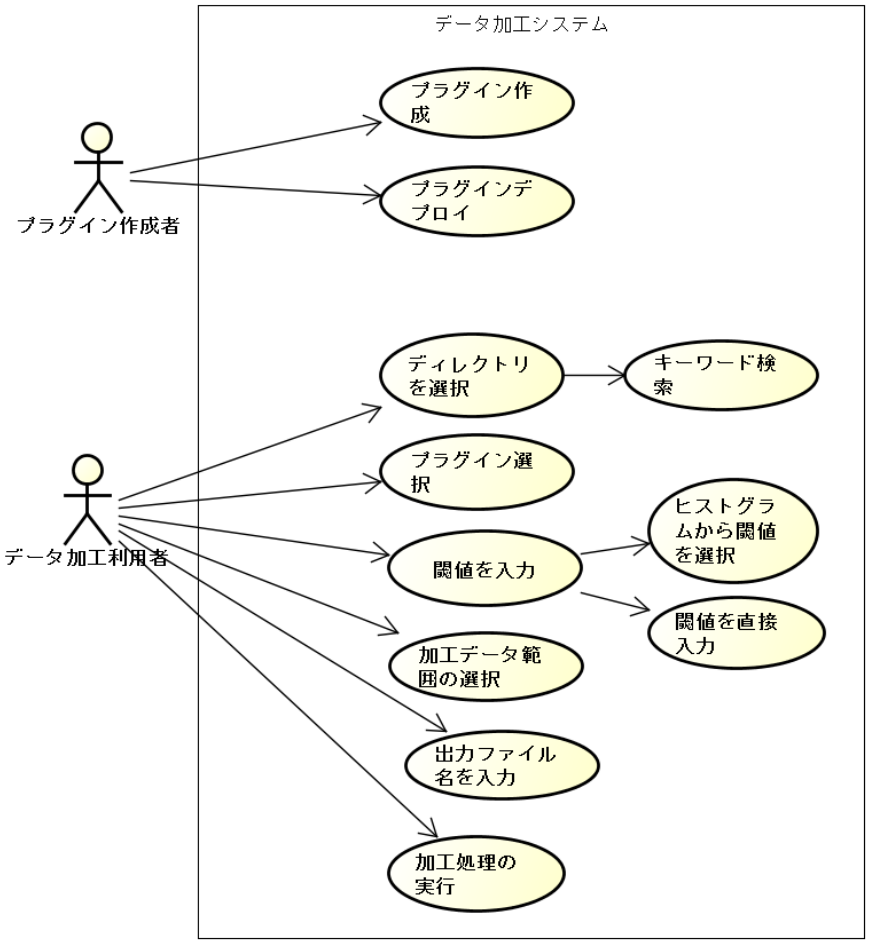


図 9　ユースケース図(データ加工システム)

5.2.3. データ分析システム

データ分析システムのユースケースを図 10 に示す.

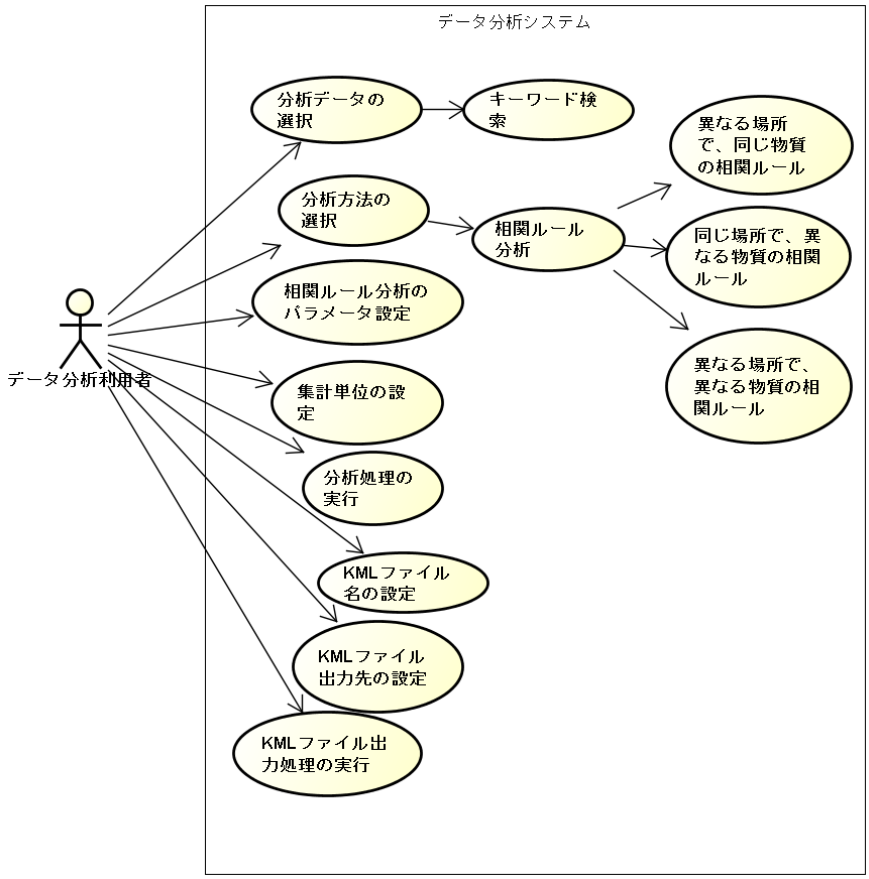


図 10　ユースケース図(データ分析システム)

5.3. プラグイン機能

本システムでは、プラグインを通して、分析対象のデータと関連するデータを取得する。その仕様を以下に示す。

【仕様 1】

分析対象のデータは横軸に属性、縦軸にログデータが格納されているデータのみ扱う

【仕様 2】

プラグインは、以下の情報を取得する。

- 分析対象データオブジェクト
- アドレスオブジェクト
- 物質属性オブジェクト
- 分析に必要な属性名とその並び順

【仕様 3】

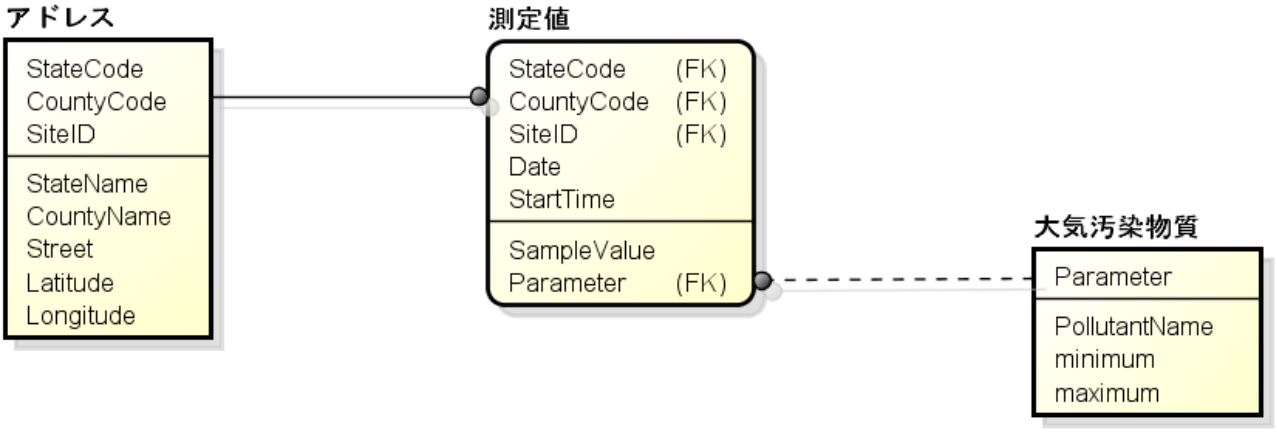
プラグインのフォルダ構成は以下のものを想定している。

Plugin フォルダー class　プラグインの class ファイルを格納

　　－ etc　　各プラグインのマスタデータを格納

【仕様 4】

プラグインの属性名と分析対象の属性名が一致しない場合は、プラグインが不一致とする



5.4. その他の機能

主な機能の補助として，以下を実装する．

5.4.1. ヒストグラム作成機能

入力データのヒストグラムを作成する．利用者はヒストグラムを参考にし，閾値を決定する．

5.4.2. スライドバーによる閾値の入力

閾値をスライドバーの移動により，閾値を入力できるようにする．

5.4.3. 中間ファイル検索機能

中間ファイル名のキーワードを入力し，該当するファイルを表示する．

5.4.4. 相関ルール分析パラメータの変更機能

相関ルール分析のパラメータを変更する．変更したパラメータを用いて，再度，相関ルール分析を行う．

5.4.5. 時間もしくは空間単位を変えての分析

中間データを時間軸もしくは空間軸で，データを集計し，分析を行う．

6. 用語集

プロジェクトに特有の用語をすべて定義する．

表 8　用語の説明

番号	用語	説明
1	相関ルール	ログデータの中から，属性間の関連を表したもの 例えば，属性 A があるならば，属性 B もあるというルールである

システム評価項目

筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻
3piece　　永田佑輔　沈雄　田中節子
2012/01/06

目次

1. はじめに

本資料はシステムの評価項目をまとめたものである。次章以降に、入力データセット、評価項目、実験環境、実験方法について述べる。

2. 入力データセット

米国環境保護庁(United States Environmental Protection Agency)が、Web 上で公開している大気汚染観測ログデータ(Air Quality System Data)を利用する。URL は http://www.epa.gov/ である。

評価には、以下の大気汚染観測ログデータを利用する。() は AQS で規定した汚染物質コードを示す。ログデータ量は合計で約 25GB である。

- Lead（12128）
- Carbon Monoxide（42101）
- Sulfur Dioxide（42401）
- Nitrogen Dioxide（42602）
- Ozone（44201）
- PM-10（81102）
- PM-2.5（88101）

システムには、以下の属性を持つデータファイルを入力する。

- SampleValue.txt（ログファイル）
  Statecode, Countycode, Siteid, Prameter, Date, Starttime, Samplevalue
- Parameter.txt（マスタファイル）
  Parameter, PollutantName, minimum, maximum
- Address.txt（マスタファイル）
  StateCode, CountyCode, SiteID, StateName, CountyName, Street, Latitude, Longitude

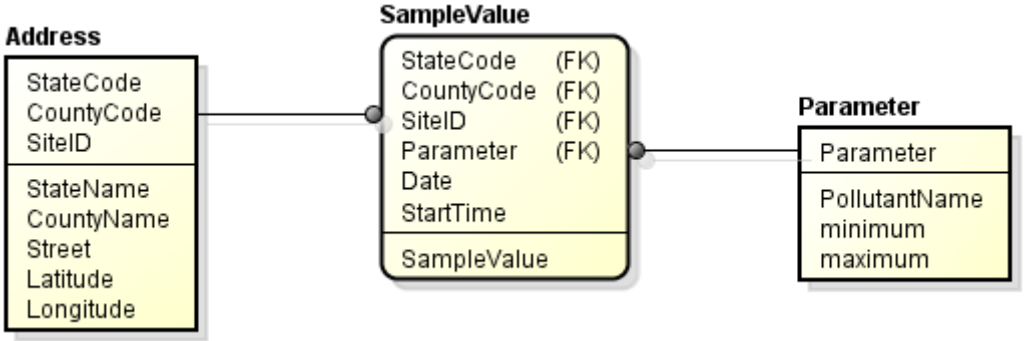


図 1　入力データの関係

本システムは、データ加工システムとデータ分析システムから構成される。

データ加工システムでは SampleValue.txt と Parameter.txt を用いて中間データファイルを作成する。

データ分析システムでは中間データファイルと Address.txt を用いて kml ファイルを出力する。

入力ファイル構成を以下に示す。ファイルの中で、各項目はタブで区切られている。

- SampleValue.txt

Name	Type (Length)	Req'd	Key	Description
State Code	CHARACTER (2)	✓	✓	State コード
County Code	CHARACTER (3)	✓	✓	County コード
Site ID	CHARACTER (4)	✓	✓	Site コード
Parameter	CHARACTER (5)	✓	✓	汚染物質コード
Date	CHARACTER (8)	✓	✓	測定年月日 （YYYYMMDD）
Start Time	CHARACTER (5)	✓	✓	測定時間 （HH:MM）
Sample Value	NUMBER (5.5)	✓		測定値

前提条件

- 一つの Parameter において、測定値の単位は同一であること

- Parameter.txt

Name	Type (Length)	Req'd	Key	Description
Parameter	CHARACTER (5)	✓	✓	汚染物質コード
PollutantName	VARCHAR2(20)	✓		汚染物質名称
minimum	NUMBER (5.5)	✓		有効最小値
maximum	NUMBER (5.5)	✓		有効最大値

- Address.txt

Name	Type (Length)	Req'd	Key	Description
State Code	CHARACTER (2)	✓	✓	State コード
County Code	CHARACTER (3)	✓	✓	County コード
Site ID	CHARACTER (4)	✓	✓	Site コード
StateName	VARCHAR2(20)	✓		State 名称
CountyName	VARCHAR2(20)	✓		County 名称
Street	VARCHAR2(50)	✓		Site 名称
Latitude	NUMBER (5.8)	✓		緯度
Longitude	NUMBER (5.8)	✓		経度

- 中間データファイル

Name	Type (Length)	Req'd	Key	Description
Date	CHARACTER (8)	✓	✓	測定年月日 （YYYYMMDD）
Start Time	CHARACTER (5)	✓	✓	測定時間 （HH:MM）
State Code	CHARACTER (2)	✓	✓	State コード
County Code	CHARACTER (3)	✓	✓	County コード
Site ID	CHARACTER (4)	✓	✓	Site コード
PollutantName	VARCHAR2(100)	✓		汚染物質名称 ※1

※1　汚染物質名称は半角ダブルコーテーションで囲む。汚染物質が一つの場合”PollutantName”となり、複数の場合” PollutantName,PollutantName”と半角カンマでつなぐ。

3. システムの評価項目

3.1. 評価パラメータの洗い出し

我々は、システムの評価を行うのにあたり、実行速度と抽出されたルールの二つの側面を考慮して評価項目を洗い出した。

- 実行速度に関して
  - データ加工システム、データ分析システムに共通
    - データサイズ
      入力データのデータサイズを変え、実行速度を測定する。
    - コンピュータの台数
      コンピュータの台数を変え、実行速度を測定する。
    - Map/Reduce の数
      Map/Reduce 数を変え、実行速度を測定する。
    - HDFS のブロックサイズ
      HDFS のブロックサイズを変え、実行速度を測定する。
  - データ加工システムに固有
    - 閾値
      閾値を変え、実行速度を測定する。
  - データ分析システムに固有
    - 相関ルール
      - ある物質に関する場所毎の相関ルール
      - ある場所に関する物質毎の相関ルール
      - 物質と場所の相関ルール

      ※各相関ルールに関して、以下のパラメータの組み合わせがある。
      - 年/月/日/時間
      - State/County/SiteID
      - 集計（100%/50%/0%超）
      - 相関ルール抽出における最小サポート値，最小確信度

- 分析内容に関して
  - GoogleEarth の表示
  - 結果の比較

3.2. 評価パラメータの選定

3.1 節の中から、優先度と評価に必要な時間を考慮し、評価パラメータは以下の項目とした。

- 実行速度に関して
  - データ加工システム、データ分析システムに共通
    - データサイズ

      システムに入力するデータサイズを 1.0/12.5/25.0 G[byte]とし、速度を比較する

      （理由）システム実行速度に与える影響が大きいと考えられるため。

  - データ加工システムに固有
    - 閾値

      各物質の閾値を、汚染値の最大値の 10 / 50 / 90 %として、速度を比較する

  - データ分析システムに固有
    - 相関ルール

      各相関ルールに関して実行速度を測定する。
      - ある物質に関する場所毎の相関ルール
      - ある場所に関する物質毎の相関ルール
      - 物質と場所の相関ルール

      各相関ルールのパラメータは以下のものとする。
      - 年/月/日/時間
      - State/County/SiteID

      以下のパラメータは固定値とする。
      - 集計=50%
      - support=0.1、conficende=0.8

- 分析内容に関して
  - GoogleEarth の表示
  - 結果の比較

4. 実験環境

システムの評価は 1 台のコンピュータの中に仮想マシン環境を構築し、ゲスト OS を三つ動作させて行う。

表 1 に、仮想マシン環境を構築するコンピュータの性能を示す。

表 1　コンピュータ環境

項目	説明
OS	Mac OS X 10.6.8
CPU	2 × 2.66 GHz Dual-Core Intel Xeon
メモリ	4 GB 667 MHz DDR2 FB-DIMM
仮想化ソフト	Virtual Box

一台の仮想マシン環境を表 2 に示す。

表 2　仮想マシン環境(1 台)

項目	説明
OS	CentOS 5.6
CPU	1 CPU
メモリ	500 MB
hadoop	hadoop-0.20.203.0
java	1.6.0_29

Hadoop 環境を表 3 に示す。

表 3　Hadoop 環境

項目	説明
Map 数	2
Reduce 数	1
HDFS のブロックサイズ	64 MB

5. 実験方法

実験手順、測定する時間を以下のように定める。

- 実験手順
  - ① Hadoop コマンドでプログラムを実行する。実行に必要なパラメータは引数で渡す
  - ② Date オブジェクトより、プログラム開始時間と終了時間を取得する
  - ③ 差分をとり、実行時間を求める

- 測定する時間
  - データ加工システム
    データ加工システムの画面上でデータ加工ボタンを押下してから、中間データを出力するまでの時間

  - データ分析システム
    データ分析システムの画面上で相関ルール表示ボタンを押下してから、画面に相関ルールを表示するまでの時間

5/12　研究開発PJ会議

議題

チーム名　次回までに考える

リーダー（渉外）　永田
副リーダー　沈
成果物管理　沈・田中　ソフトウェアで管理する
品質管理　永田

会議中の役割（ローテーション）
司会
議事録

リーダーの仕事：メンバーの代表。PJの進行は合議体で決める。
スケジュール：googleカレンダーを使う。
決まったこと：チームのメーリングリストで流す。
ミーティング：2回/週（火・木曜日　11:00～12:00最大2時間まで）
週報：進捗報告、振り返り、情報交換、今後の予定

次回　5/19（木）9:00～

5月19日　打ち合わせ

出席者：天笠先生、田中、永田、沈
議事録：田中

開発環境：
クラウド（日立のサービスを使う）
アプリ開発には、日立の開発ツールを使うかもしれない。
日立の社内資料でHadoopの勉強ができる。

どのログを分析するのかは未定
Webアクセスログ？個人のWebアクセスログ？
個人の操作ログ？

アプリケーションのイメージを決める。
全員で一つのWeb？アプリケーション
場合に応じて、日立の方と直接やり取りしてもよい。

2週間に1回集合して、進捗を天笠先生に報告する。
先生と日立の方が顧客になる。

中間報告：
まずは、hadoopで勉強したことを報告する。
開発アプリ概要
ウォーターフォールで開発が望ましい。
チケット駆動開発かもしれない。

最終報告：
開発アプリ概要
各自の担当部分

次回の打ち合わせ：

24日10時現地集合。服装は自由。
秋葉原ダイビル14階　ゼミ室

研究開発プロジェクト打合せ議事メモ（案）

日時；平成23年5月24日（火） 10:00～11:00

場所：筑波大学東京リエゾンオフィス・ゼミ室1 参加者： 土田，西澤，北川，川島，田中，永田，沈，天笠（記録）

1. 自己紹介 プロジェクトメンバーの自己紹介があった.

2. 研究開発プロジェクトのターゲットについて 研究開発プロジェクトのターゲットについて議論した. 可能性のあるターゲットは以下の通り： ・PC操作ログの分析 PCにログ取得ソフトをインストールして，ログを収集する. 事業部の業務に基づくケース分析等がある. ログ取得のためのAPIは現在も検討中であり、確認の必要がある. ・Webアクセスログの分析 PC操作ログのようなケース分析結果はなく，何を分析するかについては自分で検討する必要がある. 研究所では，以下のような取り組みがあることが紹介された： ・PC操作ログを分析し，作業効率を評価する. ・Webアクセスログを分析し，Webアプリの性能分析に利用する. また，集合 値的なアプローチで，Web UIのデザイン改善などに利用する. 日立での取り組みについて関連資料を提供していただき，それをプロジェク トメンバーで検討し，開発対象を検討することになった. また，7/1の中間 発表会に向けて，Skypeを利用して密に議論を行うことになった. 5/31（火），6/7（火），15（水），22（水） ※いずれも19:00から. Skypeのアカウントは，後日メールで回覧.

3. マイルストーン 今後の主なマイルストーンは以下の通りである. 7月1日（金）中間発表会1（プロジェクトのターゲット像を報告） 10月7日（金），14日（金）中間発表会2（開発状況等の報告，デモ） （翌年）1月第3週 修士論文・特定課題研究報告書提出締切 （翌年）2月第1週 修士論文・特定課題研究発表会

4. Hadoop，ストリームに関する講習について Hadoopおよびストリーム（Flexによる可視化）に関する講習をプロジェクト 向けに開講して頂けるか検討していただく. ・それぞれについて半日程度. ・演習環境が準備できない場合は座学中心に. 6/29（水）の日立uCSDPアカデミック支援技術交流会に合わせて開講？

5. 開発環境について 以下のマシンを天笠が提供する： Mac Pro（4x Xeon，4GB memory，2x 500GB disk）は提供可能.

6. その他 開発モデルは基本的にウォーターフォールモデル. 開発ドキュメントは，PBLの資料をベースに作成. 研究開発プロジェクト議事録

●日付：2011/05/26　11:30～12:20
●場所：3F805

●作成者：永田佑輔

●参加者：永田、田中、沈

●議題
　◎インストールするソフトウェアの決定
　◎チーム名の決定

●会議内容と決定事項：

　◎Virtual Box上で動く開発環境の構築について(田中・沈)
　　・インストールするOS

CentOS 最新バージョン
仮想環境の構成はメモリ1ＧＢで、パーティションはデフォルト設定にする
・インストールするソフト
Hadoop 安定版で最新バージョン

◎Macにインストールするソフトウェア
・VNCでリモートデスクトップ環境を構築する（沈）
・SVN(subversion)　バージョン管理(document source)（永田）
・Virtual Box（沈）
・Redmine（永田）

◎Macと仮想マシンのrootパスワード
sendouteki

◎チーム名
3 piece

◎メーリングリスト
メーリングリストは以下の３通りを考えておく
・開発メンバーのみ
・開発メンバーと天笠先生
・開発メンバーと天笠先生と日立二人

◎クライアント環境(導入は各自)
・Java　java6.0 update 25
・eclipse IDE for Java Developers
・VNC

●アクション項目：
なし

●次回の予定

研究開発プロジェクト議事録

日付：2011/05/31　19:10～20:00
場所：インターネット(Skypeでのミーティングのため)

作成者：永田佑輔

参加者：
土田さん、永田、田中、沈

1.内容の要旨：
・Hitachi IT Operations Directorについて質疑応答
・Hadoopの学習と開発環境の理解

2.詳細内容：

■Directorについて(土田さん)
・北米の開発部隊が作った製品、日本発の製品ではない
・アクセスログの管理と運用管理ツールである
個人用ソフトや正式なライセンスではソフトがインストールされて

いないかチェックする。また、情報流出を検出する。

■Analyzerについて(土田さん)
・PCの中で、ログを検出して、ルートコーズアナリシスが行える機能を
持っている
・エラーログを元に、技術的な解決を表示する
・Directorと相互運用可能である

■電子情報通信学会の資料について
・電子情報通信学会の資料が有料で見れない(田中)
→　土田さんが共著者である西澤さんに資料をいただけるか伺う

■ログのフォーマットについて
・ログのフォーマットはどのような形式ですか？(永田)
・ログのフォーマットと格納場所はプラットフォームに依存する(土田さん)
・どういうことをやりたいのかを決めないと、どのログを分析していいか
わからない(土田さん)

■DirectorのHadoopの使用について
・DirectorはHadoopなどを利用していますか？(沈さん)
→　DirectorはHadoopは使用しておらず、ログ収集を行っています。(土田さん)

■Hadoopの環境について
・Hadoopは入力としてHDFSを利用すると考えていいのか？(田中)
→　Hadoopへの入力として、HDFSやRDBなど様々なものが考えられる。
しかし、現状では入力としてどのようなデータを扱うのか明確に
なっていない。したがって、今のところは、HDFSを入力として、
MapReduce関数により、並列処理を行えるようにする。(土田さん)

■Hadoopの学習について(土田さん)
・日立製作所様がHadoopの講習会を開いて頂ける
学生は事前にHadoopでしたいことやわからないことをまとめておく
・HDFSにデータが入っていて、MapReduceで処理をする学習をしておくと良い
・データの入出力にストリームができるので、講習会で教える

■次回までのプロジェクトの進め方について
1.いただいた資料を分割し、それぞれに担当を分ける
2.担当者は、担当したところをレポートにまとめ、チームで共有する
3.実現したいことを３個リストアップする
このとき、調査として、アクセスログだけではなくて、広く考える。
マーケティングをする。また、どのログをとるのかを考える。

■会議の進め方について
・毎回、議事録を取り、議論とアクションアイテムをまとめる
・会議の始めに、前回の会議のアクションアイテムの進捗を確認する

3.次回までの活動

学生：
・いただいた資料を元に、チームで行いたいことを３個考える
・考えた案の実現性を考える
・考えた内容をドキュメント化し、次回の参加者で共有する

土田さん：
・シンポジウムでのスライドを学生に送る
・「業務モニタリングへ向けた映像ログからの人物適応動作解析」の
資料を西澤さんからいただけるか確認を取る

4.次回の予定
6月7日(火)　19時から～

研究開発プロジェクト議事録

日付：2011/06/02　11:00～11:30
場所：3F805室

作成者：沈

参加者：
田中、沈

目的：Hadoopを用いた大規模ログ分析システムの提案

田中の提案：
テーマ：学内Web検索のkey-wordのランキング
目的：筑波大学の学生の学習状況の分析
概要：各PCにソフトウェアを設定して、各browserで検索Logを開発サーバに送信して、サーバは収集した情報を分析して、結果は画面で表示する

沈の提案　：
テーマ：英語文章の解析システム
目的：英語勉強の場合、単語の頻度の分析
概要：利用者は分析したい文章を登録して、分析した結果をフィードバックする。

次回の予定
6月9日(木)　11時から～　　(沈　欠席)
研究開発プロジェクト議事録

●日付：2011/06/07　19:00～19:50
●場所：インターネット(Skypeを利用したミーティング)

●作成者：田中節子

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、田中、沈

●フォローアップ案件：
◎学生提案に関するフィードバック

沈、田中の提案に共通
・研究開発プロジェクトのテーマとしては容易である。
・Hadoopを利用して分析するには入力データが少ない。

・このシステムを利用することによって、利用者にどういうメリットがあるのか明確でない。
・入力データの分析結果を表示するだけでは、システム化する意味が無い。

望ましいシステム
・利用者が欲しくなり、利用者の役に立つシステムを考える。
・利用者がワクワクするシステムを考える。

●アクション項目：
◎基礎的な知識が不足しているため、典型的なWebアクセス分析について学ぶ
・天笠先生から教えていただいたURL等を参考にして調べる。
・期限は2011年6月9日（木）

◎メンバー全員でブレインストーミングを行う
・一人につき10～20個の案を出して話し合う。
・議論は収束しても、発散しても良い。
・先生方に、入力データが入手可能か伺ってもよい。
・期限は2011年6月9日（木）

●決定事項
◎ブレインストーミングした結果を、関係する皆様にメールで送信する。
・期限は2011年6月9日（木）の夜
研究開発プロジェクト議事録

●日付：2011/06/21　22:00～22:15
●場所：インターネット(Skypeを利用したミーティング)

●作成者：田中節子

●参加者：
永田、田中、沈

●議題
研究開発プロジェクトのテーマの検討

●決定事項
◎6/22の15:00からskypeミーティング
◎各自、ログデータの入手方法、hadoopを使って何を分析するかを用意しておくこと

●議論内容

◎やはりログデータを用いてhadoopで分析することになった。
→どのデータをどのように分析するべきか？を議論した。

◎沈さんの案
参考ＵＲＬ
http://www.diva-gis.org/climate

http://www.nathankerr.com/projects/parallel-gis-processing/gis_on_hadoop.html
http://www.hadooper.cn/dct/attach/Y2xiOmNsYjpwZGY6MTEy

研究開発プロジェクト議事録

●日付：2011/06/22　19:30～21:30

●場所：インターネット(Skypeを利用したミーティング)

●作成者：田中節子

●参加者：
土田様（日立）、西澤様（日立）、永田、田中、沈

●議題
◎研究開発プロジェクトのテーマ策定

●フォローアップ案件：
◎学生提案に関するフィードバック

田中案
・日本全国の大気汚染状況の分析
・全国の大気汚染状況を図で可視化する。
・地域毎に比較できるようにする。（異なる都道府県同士、同一都道府県内の各地域）
・時系列で比較できるようにする。
・（できれば）大気汚染と健康の関係を示す。
・参考URL
http://www.nies.go.jp/igreen/index.html
http://soramame.taiki.go.jp/DataMap.php
http://www.kankyo.metro.tokyo.jp/air/air_pollution/result_measurement.html
http://www.env.go.jp/kijun/taiki.html
先生方からのフィードバック
・このシステムを何のために作るのか明確にする。
・メッシュ、時間をどうのように区切ってどのように分析するのか明確にする。
・粒度を変えることによって計算量も変わるので、hadoopを利用する価値はある。
・大気汚染データの可視化の価値を明確にする。

沈案
・hadoopでGISデータを解析するCO2濃度、雨量の分析、土地利用動向の分析
（GISデータを選択理由は：データ標準格式がある、ネットでダウンロードして集めて、大きいサイズが入手できる）
・自然環境条件図、災害履歴図、土地利用動向図を作成する
・参考URL

https://docs.google.com/document/d/17dUwCkzXb-aUun7UJ-YBDD8Gp8vqRLb-EHSLuuf35uU/edit
先生方からのフィードバック
・このシステムを何のために作るのか明確にする。

永田案
・JP1の機能をhadoopで実現できないか
先生方からのフィードバック
・JP1とhadoopはあまり関係ない。

●アクション項目：
◎学生3人で開発テーマを一つに決める。
・開発テーマをpptスライド２、３枚で表現する。
・システムのねらい、目的、システム開発の構成を書く。
・締め切りは6月23日（木）

●決定事項
◎天笠先生に今日の会議内容についてメールで報告する。
・締め切りは6月22日（水）

◎議事録とスライドを一緒に、土田様、西澤様にメールで送信する。
・締め切りは6月23日（木）

◎議事録とスライドを天笠先生にメールで送信する。
・土田様、西澤様にスライドのコメントをいただいた後、修正してから送信する。

●今後の予定
◎Hadoopの講習会を6月29日(水)の10時から行う。
研究開発プロジェクト議事録

●日付：2011/07/07　11:20～12:20

●場所：インターネット(Skypeを利用したミーティング)

●作成者：沈雄

●参加者：
永田、田中、沈

●議題
◎分析データと可視化を検討する

大気汚染データのURL　http://soramame.taiki.go.jp/DataMap.php
国土数値情報データ http://nlftp.mlit.go.jp/ksj/index.html
理由　各時間と地域を絞る
問題　1,ソースにデータはない、取得は難しい。
2,そのデータ（地名）→GISデータ（経緯度）に変換
3,何をユーザさんに提供する　座標　→[google earth]
その後、集計した結果→時間と場所を選べる、データを表示

●アクション項目：
1,大気汚染データの収集する方法（田中）
2,GISデータの構成（永田）
3,[google earth] APIの調査（沈）
●決定事項

●今後の予定
◎7月14日(木)の11時から行う。
研究開発プロジェクト議事録

●日付：2011/07/14　19:30～21:30
●場所：インターネット(Skypeを利用したミーティング)

●作成者：田中節子

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、永田、田中、沈

●フォローアップ案件：

◎中間報告会で指摘された事項について

①チームとしてまとまっていないのではないか
１学期は就職活動が忙しかったため、メンバーで十分に議論することができなかった。
現在は、メンバー全員で会議を行っている。

また、チームで行う会議の他に、以下の文書を残す。
・作業報告書
記入項目例：作業の名前、作業内容(できたこと、できなかったこと)、進捗率、作業見積もり時間、作業時間、締め切り、課題
・懸案リスト
記入項目：次回までに誰が何を担当するのか

② システムの利用者は誰か
大気汚染情報（ログ）を分析する必要のある人（エンドユーザが利用者ではない）。
今後チームで以下の項目を定義する。
・ユースケース図のアクター
・システムを誰がどういう立場で使うのか

③ プロジェクトの目標は何か
大気汚染情報を分析したい人に、対話式の分析システムを提供する。
大気汚染情報（ログ）に関して、時間的・空間的にさまざまな粒度で分析し、その結果をGoogleEarthを利用して可視化する。

◎システム概要について
どうやってデータを取得し、どういう分析をするのかを明確にする。
時間軸について「分析する日にちを選ぶ」等、明確にする。
空間軸について「点で表現する」等、どのくらいの範囲で表現するのか明確にする。
Hadoopで分析する仕様は、始めはシンプルなものにする。(収集したデータを利用して、簡単なデモを作成する)

●アクション項目：
◎大気汚染情報はアメリカのデータを利用する。
・既にログ形式になっているデータをダウンロードして利用する
・データを探す期限は、7月15日（金）

●決定事項：
◎土田様、西澤様、天笠先生に参加していただく全体会議を週一回行う。
・毎週水曜日19:00開始
・次回の全体会議は7月20日（水）19:00

◎先生方にシステムに関する意見をいただく場合、具体例を挙げて説明する。

研究開発プロジェクト議事録

●日付：2011/07/20　19:00～19:50
●場所：インターネット(Skypeを利用したミーティング)

●作成者：永田佑輔

●参加者：
土田様（日立）、西澤様（日立）、永田、田中、沈

●会議内容：

◎進捗報告書に基づいて報告を行った
〇使用するデータについて
AQS Data Martのデータは1980年から現在までデータがあるが、取得が困難である。したがって、同じAQSのホームページに同様のデータが1993～2011年の範囲で公開されているので、そのデータを用いる。

〇データの素性に関しては、マニュアルを参照し、理解済みである。

●アクション項目：
◎利用者と利用シーン、システムの機能を明確にし、関係者に提出する
まず、利用者と利用シーンを決める。そして、システムの機能と利用者のメリットを考える。

【期限】7月24日(日)

◎スケジュールを作成し、開発メンバで合意したものを関係者に提出する
なお、主にマイルストーンを重点的に記述する

【期限】7月24日(日)

●決定事項：
◎ドキュメントは、当日の朝までには配付し、会議までに一度目を通しておく。

◎作業内容は成果物に重点をおき、具体的に書く。また、全体の進捗度を把握するために、行っていないタスクに関しても記述する。

◎週に一回の全体会議では、主に進捗報告を行う
会議は、学生が「今週の作業内容と進捗、次週の作業方針」を述べ、助言をいただく形で行う。

◎開発方針として、要求仕様を決めてから、設計／開発を行う。

●次回の予定
7月27日(水) 19:00開始

研究開発プロジェクト議事録

●日付：2011/07/27　19:10～20:30
●場所：インターネット(Skypeを利用したミーティング)

●作成者：永田佑輔

●参加者：
土田様（日立）、永田、田中、沈

●会議内容と決定事項：

◎システム要件
環境コンサルタントを対象にするのはわかるが、環境汚染データをどう分析するのかわからない。分析内容とシステム化のメリットを調査し、ドキュメント化する。このとき、2、3個程度ピックアップする。

◎AQSデータ
データの詳細がわからない。報告したデータソースの中で、使用するデータと使用しないデータをわける。また、使用するデータの種別、項目、データ量などを明確にする。そして、これらのことをドキュメント化する。

◎開発環境の構築
・Mac Pro上で仮想マシンを動かし、開発環境を構築しているが、解決方法がわからないエラーが発生した。現在調査中だが、同様の環境をWindowsで構築したが、上記のエラーは発生しなかった。
・作業内容と発生したエラーをドキュメント化して残すこと。

◎開発スケジュール
・プロジェクトの工夫点を明確化しておく
報告した開発スケジュールだと、他の工程と比較して、要件定義工程が長いので、報告会で質問されると思われる。

・調査内容は今後使うかも知れないのでドキュメントする
要件定義工程での調査だけでなく、開発時に行った技術調査についてもまとめる

●アクション項目：

◎利用者像を明確化するために、環境コンサルタントについて以下の項目をドキュメント化する。

・仕事内容
・分析方法
・システム化による効果

【田中担当、期限：2011/07/29】

◎開発環境の構築時に発生したエラーについて、作業内容とエラーメッセージ、解決するために行った調査をドキュメント化する

【田中担当、期限：未定】

◎使用するデータの洗い出し
・ファイル容量
・レコード数
・データ種別

環境コンサルタントの項目が終了次第、担当と期限を決定する

◎state code, couty code, site idして、ダウンロード、ファイルの容量とレコード数を調査し、ドキュメント化する。
データソース:http://www.epa.gov/aqspubl1/site.html

【沈担当、期限：2011/07/29】

◎プロジェクト規約の作成

【永田担当、期限：2011/07/29】

●次回の予定
8月3日(水) 19:00開始

研究開発プロジェクト議事録

●日付：2011/08/03　19:34～20:27
●場所：インターネット(Skypeを利用したミーティング)

●作成者：沈

●参加者：
土田様（日立）、天笠先生、永田、田中、沈

●会議内容と決定事項：

◎既存の環境コンサルタント業務についての調査（田中）
コメント
土田さん　ドキュメント化が足りない（リンクの部分の説明は必要）
天笠先生　要約でリンクを説明する。

◎state code, couty code, site idのファイルの容量とレコード数の調査（沈）
コメント
土田さん state code, couty code, site idのなんのために？
田中さん　可視化のために、「google earth」を渡せるにように。

●アクション項目：
◎毎日コアタイム。（朝10時～11時）

◎アメリカのデータの分析方法の調査

【田中、沈　担当、期限：月曜日の正午】

◎コンサルタント業務（目的など）

【永田担当、期限：月曜日の正午】

●次回の予定
8月12日(金) 19:00開始

研究開発プロジェクト議事録

●日付：2011/08/12　19:30～20:40
●場所：インターネット(Skypeを利用したミーティング)

●作成者：永田佑輔

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、永田、田中、沈

●会議内容と決定事項：

◎日本の大気汚染分析
・6～9時間平均は、6と9は含んでいるのか不明
・平均と分散を求めるだけでは分析になっていない
平均と分散は現状を知るために使う

◎アメリカの大気汚染分析

■分析データの使われ方
・Google Mapを使う場合
位置情報(緯度と経度)を元に分析結果をgoogle earthで可視化する
場所をクリックすると、濃度を折れ線グラフで表示する

・Google Earthを使わない場合
一日を折れ線グラフで表示して、過去と現在を比較する

■扱うデータについて
dailyは一日、それ以外は一時間ごとに使用する

位置情報は提供されているの？
http://www.epa.gov/aqspubl1/site.html

◎環境コンサルタントについて
環境コンサルタントは国が示している基準を元に計算し、国や地方自治体に提供している

◎データ分析案
相関ルールマイニング
詳細は「相関ルールマイニング」、「時系列」、「DTW」を調べる

●アクション項目：

◎データマイニングの勉強
【担当者：永田、田中、沈】
【期限：8月14日(日)】

◎土地や気温と大気汚染データの関連を調査する
【担当者：永田、田中、沈】
【期限：8月16日(火) 22時】

◎コンサルタントに分析方法を聞く
【担当者：土田様】

●次回の予定
8月17日(水) 19:00開始

研究開発プロジェクト議事録

●日付：2011/08/17 19:00～20:00
●場所：インターネット(Skypeを利用したミーティング)

●作成者：田中節子

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、永田、田中

●フォローアップ案件：

◎データマイニングの勉強
（沈）時系列データマイニング（決定木によるデータマイニング、ヒントとしてのユーザ入力）、DTW によるマッチング
（永田）Web記事を参照し習得。
（田中）相関ルールマイニング、時系列データマイニング、ダイナミックタイムワーピング距離

◎土地や気温と大気汚染データの関連を調査
米国の気象データを調査した。
→大気汚染データ同士の相関を取ることになった。

◎環境情報ソリューションについて（日立様からいただいた資料）
・どういう課題の下で、どういう分析結果が有効なのかが明確に書かれている。
システム化の必要性とメリットが明確に書かれているため、全体的に参考になるが、研究開発プロジェクトには特に６章が参考になる。
・この資料に書かれていた他の内容。
天候デリバティブを用いた予測システム。

　　天候デリバティブとは、気温、降雨量、降雪量、湿度、台風、波の高さ、落雷、降霜などがあらかじめ決めた基準値を超えた場合に、補償金を受けられる金融派生商品のこと。
　　予測に偏微分方程式を利用する。

●アクション項目：
◎相関ルールのプログラムのプロトタイプを作成する。
　存在すればHadoopの分析パッケージ、又は、Rを用いる。
　相関ルールとは、大量のデータから、頻繁に同時に生起する事象同士を相関の強い事象の関係、すなわち相関ルールとして抽出すること。
　締め切りは、8月23日。

◎システム化のメリットを明確にする。
　システム化の必要性を明確すれば、メリットも明確になる。
　締め切りは、8月23日。

◎相関ルールを抽出する対象を明確にする。
　締め切りは、8月23日。

◎大気汚染の一時間当たりの測定値を、一日当たりの測定値に変換するプログラムを書く。
　→チーム内で話した結果、これは当面見送ることになりました。

●決定事項：
　次回の全体会議は8月24日（水）20:30
研究開発プロジェクト議事録

●日付：2011/08/24　20:30～22:00
●場所：インターネット(Skypeを利用したミーティング)

●作成者：永田佑輔

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、永田、田中、沈

●会議内容と決定事項：

◎相関ルールを抽出する対象を明確にする(沈担当)
・資料は平均値と標準偏差の関係を調べたものである
・資料の図と下部のデータは関連性はない
　図は説明のため準備したものである
・これを平均と標準偏差を使った理由は？
　異なる物質の相関を求めるために、平均と標準偏差を使った
・事象を洗い出して、どの事象を検討するかが大切である
・数値データの分布から、閾値を求めて、二値化する。
　会議では、この二値化について、議論を行いたい。
　(汚染されているか、されていないかの閾値を議論したい)
　なお、一つの物質を一つの現象として、それぞれの閾値を求める。

◎相関ルールのプログラムのプロトタイプを作成する(田中担当)
・Hadoopの相関を求めるパッケージが見つけられなかったので、
　Ｒを用いて相関ルールを求めた
・相関係数はspearmanで求めた。pearsonは正規分布になっている場合はよいが、正規分布になっていない場合は精度が悪い
・相関ルールと相関係数は違う

◎田中担当分と沈担当分の違いは
・沈の担当は異なる物質を比較して、相関ルールを見つけること
・田中の担当はプロトタイプを作成し、沈が求めた相関ルールを適用すること

◎システム化のメリットを明確にする(永田担当)
・土田さんのヒアリング結果でメリットが明確にならない場合の
対策が決まっていない
・担当者だけでは明確にするのが難しいので、このタスクは日立側が担当する

●アクション項目：

◎各物質の汚染データの分布を見て、閾値を決め、二値化する
その結果を天笠先生にメールする
【締切：8/29】

担当は以下のとおりである。担当は物質名で記述する。

〇永田分
Lead - PB - daily(12128)
Carbon Monoxide(42101)
Sulfur Dioxide(42401)

〇田中分
Nitrogen Dioxide(42602)
Ozone(44201)

〇沈分
PM 10(81102)
PM 2.5(88101)

●次回の予定
8月31日(水) 19:00開始

研究開発プロジェクト議事録

●日付：2011/08/31　19:30～20:30
●場所：インターネット(Skypeを利用したミーティング)

●作成者：田中節子

●参加者：
土田様（日立）、永田、沈、田中

●フォローアップ案件：

◎相関ルールを抽出するために、各汚染物質のしきい値（中央値）を求める
永田担当
Lead - PB - daily(12128)
Carbon Monoxide(42101)
Sulfur Dioxide(42401)

田中担当
Nitrogen Dioxide(42602)
Ozone(44201)

沈担当

PM 10(81102)
PM 2.5(88101)

いただいたアドバイス
データ分布の状況から、しきい値は中央値ではない方が相関ルールを抽出しやすい。
上位10%、 20%、 30%...と変えて、どの辺りが的確に相関ルールを抽出できるか試す。

●アクション項目：
◎データのクリーニングをする。
現在の測定値には、マイナス値も含まれている。
各汚染物質について、どこからどこまでを有効値とするのか明確にする。
締め切りは、9月1日。

◎各汚染物質について、しきい値を求める。
しきい値を上位10%、 20%、 30%...と変えて、どの辺りが的確に相関ルールを抽出できるか試す。
担当は上記と同じ。
締め切りは、9月7日。

●決定事項：
次回の全体会議は9月7日（水）19:00
研究開発プロジェクト議事録

●日付：2011/09/07　19:34～20:40
●場所：インターネット(Skypeを利用したミーティング)

●作成者：永田佑輔

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、永田、沈、田中

●会議内容：

アクション項目の結果

◎データのクリーニングをする。
現在の測定値には、マイナス値も含まれている。
各汚染物質について、どこからどこまでを有効値とするのか明確にする。

【結果】
汚染値がマイナスの値は異常値なので、除去する。汚染値が0のものは測定器の
精度により、存在するので、除去しない。よって、有効値は0以上とする。

◎各汚染物質について、しきい値を求める。
しきい値を上位10%、 20%、 30%...と変えて、どの辺りが的確に相関ルールを抽出できるか試す。

【結果】

〇田中の分析方法
バスケットデータへの変更方法は、まず、閾値を元に数値を二値化する。そして、
横軸に地域ごとの値を、縦軸に時間を取った行列を作成する。このデータを用いて
相関ルールを抽出する。

【参加者からの意見】
・上位10%／20%を元データから除去した値から、相関ルールを求めるのではなく、
上位10%／20%に該当する値から、求めること

〇沈の分析方法
Rのapriori関数を使って、相関ルールを求めた

【参加者からの意見】
・1998年だけでは、12月に汚染度が高いとはいえない

〇分析方法について
・チームで分析方法は決めていない

アクション項目全体を通して、参加者からの意見
・役割を物質で分担するのは、あまり効率がよくない
物質が違うだけで、やってることは同じ
・チーム会議の議事録を残し、プロジェクトの記録を終えるようにすること
この記録が発表資料作成や修士論文の内容となる

10月に行う中間報告について
10月の報告会では、完成品を作る必要はない。完成品の一部を作り、
デモを行う

相関ルールを求めるための分析対象
相関ルールを求める対象は以下の3通りが考えられる
・同一時刻で異なる地域の汚染値
・異なる時刻で同一物質の汚染値
・異なる物質での汚染値

●アクション項目：

◎相関ルールの抽出
上記3通りの相関ルールをどのようにして抽出するのか決める
【担当者：全員】
【期限：9/13】

◎システム開発環境の検討
実際のシステムでは、どこでHadoopを使うのか、Rを統計ツールとして利用するのか、等を明確にする。
【担当者：全員】
【期限：9/13】

●決定事項：

◎定例会議の議事録を残す

●次回の予定
次回は9月14日(水)19時から

研究開発プロジェクト議事録

●日付：2011/09/22　19:15～20:00
●場所：インターネット(Skypeを利用したミーティング)

●作成者：永田佑輔

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、永田、沈、田中

●会議内容：

アクション項目の結果

◎異なる時刻で同一物質の相関ルール抽出
今回は全米で出したが、日にち単位で測定場所ごとに相関を取る

【参加者からの意見】
■出てきた相関ルールの意味は？
→　因果関係を表している
→　相関ルールは因果関係を求めるものではない

◎システムの要件、利用者を確認する。
配布した資料を元に、説明を行った

【参加者からの意見】
■システムを作れる段階ではない
・データ処理の流れがない
・データ加工の部分も、どういう風に加工していくのかが書いてない
システムが構成するモジュールの流れを書くこと
また、ユースケース図やアクタが何かを明確化する

●アクション項目：

◎システムが開発できるように資料を修正する
具体的には以下の作業を行う
・モジュール間のデータの流れを書く
・データ加工部を詳細化する
・データ分析部を詳細化する

◎システムのユースケース図の作成とアクタの明確化

◎画面の操作に対応したシステムの処理を明確化する

●決定事項：

◎分析方法について、これまでの結果を記録しておく
具体的には、パラメータとその結果を記録する

◎分析時は、時間軸と空間軸の切り替えが行えるようにする
時間軸：年／月／日
空間軸：州／郡／サイト
例えば、月単位で、郡ごとに分析を行う

●次回の予定
次回は9月28日(水) 19時から

研究開発プロジェクト議事録

●日付：2011/10/05　19:00～20:30

●場所：インターネット(Skypeを利用したミーティング)

●作成者：田中節子

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、永田、沈、田中

●フォローアップ案件：

◎進捗報告
開発システムの説明
・システムの概要
・ユースケース図
・アクターの明確化
-プラグイン作成者
-利用者
・クラス図
・シーケンス図

◎先生方からいただいたコメント
・利用者が見てシステムを理解するドキュメントを書く。
・KMLの出力イメージを具体的にする。
・システムの実現可能性はあるのか明確にする。
・開発とドキュメントは並行して行う。
・10/14のデモでは、入力画面、入力データ、中間データファイル、出力のイメージ、システムの処理の流れを明確にする。

●アクション項目：
◎利用者が見てシステムを理解するドキュメントを書く。
締め切りは、10月19日。

●決定事項：
次回の全体会議は10月19日（水）19:00
研究開発プロジェクト議事録

●日付：2011/10/21　19:00～20:30
●場所：インターネット(Skypeを利用したミーティング)

●作成者：沈

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、永田、田中、沈

●フォローアップ案件：

◎報告会について、先生と日立方からいただいたコメント

１、土田さんの指摘：何々さんの指摘を明確する。

２、ゴール（目的）は不明について
このシステムは大規模ログをHadoopを利用して分析し、新しい価値を発見をするものである。
今回は大気汚染データを利用する。
ユーザは特定しない。このシステムの場合、環境コンサルタント等を仮定する。

天笠先生、土田様、西澤様の思いと意識合わせた。

３、西澤さんの指摘：Rをなのためにおよび使う方を説明する。

４、プレゼンの反省、メンバー間と意識合わせる。最終報告書は日立さんにコメントをいただく。

●アクション項目：
◎開発構想書の作成
詳しくスケジュールの再制定。
設計に関する図の作成

●決定事項：
次回の全体会議は10月26日（水）19:00
研究開発プロジェクト議事録

●日付：2011/10/26　19:45～21:45
●場所：インターネット(Skypeを利用したミーティング)

●作成者：永田佑輔

●参加者：
西澤様（日立）、天笠先生、永田、沈、田中

●会議内容：
本会議では、スケジュールと開発構想書について、意見をいただいた。
いただいた意見を以下に示す。

【スケジュールについて】
〇タスクの担当者と必要な日数を書かくこと
〇プログラムの規模をstepで表すと、実装工程の期間で実装可能か議論できる

【開発構想書について】
◎1.2.1 開発の目的
〇時系列データはシーケンスに依存しているので、単純に分割できるとはいえない
説明としては、相関ルールの分析手法を述べ、その分析手法ではこういう分割ができるということを述べる
〇Hadoopの有効性を示すわけではなく、開発するのが目的である
大規模ログ分析システムを作ると述べ、その後に、扱うデータを述べる。
扱うデータは、一例として、大気汚染ログデータを扱う
〇時間がかかる等の定性的な表現が多い。定量的に書かれていると、わかりやすい。
事例を踏まえて、数値を使って書くと良い。
〇あいまいな表現がある。作文は厳密に書くこと。
〇本システムを使った時の、目標とする実行速度とその理由を書くこと
従来は、相関ルールを見つけるのにどのくらいの時間がかかっているのか？
〇プロジェクトの目標を明確にすること
以下のどちらを言いたいのか、チームで議論すること
・一つのジョブが数時間かかっていたが、本システムを利用すると、数分になったので、実用的になった【チームとしてはこっち選択】
・一つのジョブの実行速度が何倍になったから、よかった

◎1.2.2 システムの機能
〇プラグイン形式で、「様々な」入力データの格納形式に対応するとあるが、
対応する形式を具体的に書くこと
〇相関ルールの定義を書く
itemとitemセットを述べ、その後、相関ルールの説明を行う

○具体的かつ形式的に書く
形式的にはフォーマルな書き方
○同じ意味の言葉は統一する
○曖昧な表現は具体的に書く
○1.2節と3章で同じタイトルが使われており、違いがわからない
○プラグインとは、データがソースにより変わるので、共通の関数でデータに
アクセスするためのインターフェースである
○プラグイン作成者はデータソースを読み込むための、関数を作成する
○プラグイン機能は、データ加工機能とデータ分析機能と同格ではない
プラグイン機能は、データ加工機能の一部である

◎3 システム概要
○全体図に、プラグインを書かくこと
データ加工システムの前に、プラグインがある。
○閾値を対話的に、決めることがわかるようにする
○利用者の操作と対応するシステムの処理の流れを明確にする
これは、外部設計書に書く

◎4 機能要件
○表の上の文章が説明になっていない
○表に項番をつける
○プラグインは、classファイルにして、指定したフォルダに置く
○「プラグイン」から「データ入力プラグイン」に名前を変える
○アプリオリのパラメータ設定の機能が抜けている
○中間データファイルの命名規則は工夫が必要である。中間データファイルが
増えても、一意に決定できるようにする
○Hadoopのノード数やMap数、Reduce数は、システムでは設定しない
システムの外で、設定する
○利用者に対する機能とシステムの出力が混ざってて、わかりにくい

◎5 非機能要件
○運用面に関しては、仕様であるので、機能要件に書く

◎6 ドキュメント要求
○プロジェクトとして、作成しなければいけないドキュメントはない

●アクション項目：
◎スケジュールの担当と必要な日数を書く
◎指摘していただいた意見を元に、開発構想書を修正する

●決定事項：
特になし

●次回の予定
次回は11月2日(水) 20時30分から

研究開発プロジェクト議事録

●日付：2011/11/09　20:30～21:30
●場所：インターネット(Skypeを利用したミーティング)

●作成者：沈

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、永田、田中、沈

●フォローアップ案件：

◎前回の修正した部分の検討

◎画面遷移内部処理
内部設計書に書く。

◎リスクの対策
プロジェクトのリスクを識別し、分析し、リスクに対応するための対策は必要です。

◎データの形式
内部設計書に書く。

◎開発構成書の実行時間の測定
相関ルールについて、どなん時間の測定は要らない。
しかし、報告書に事例を持ち、詳しく説明する。

◎ヒストグラムについて
ヒストグラムの表示は５分間ぐらいかかる。

●アクション項目：
◎今回いただいたご意見に基づき、各ドキュメントを修正する。
スケジュール、開発構想書、画面定義書、画面遷移を修正する。

●決定事項：
次回の全体会議は11月26日（水）19:00
研究開発プロジェクト議事録

●日付：2011/11/02　20:45～22:00
●場所：インターネット(Skypeを利用したミーティング)

●作成者：田中節子

●参加者：
西澤様（日立）、天笠先生、永田、沈、田中

●会議内容：

◎スケジュールの明確化
最優先でやるべき事である。
スケジューリング設計の目安
0.5kstep/1人月
→現在作業を細分化している。優先度を挙げて取り組む。

【画面遷移図について】

◎画面構成
Web画面の操作は一般的に、左から右に、上から下に操作する。

◎ヒストグラム
閾値を設定する際、閾値以上の値、以下の値のどちらをバスケットに入れるのかユーザに指定してもらう。
閾値を設定すると何%のデータがバスケットに入るのかを表示する。

◎中間データファイル
メタデータは入れるのか、データ型、精度等、スキーマ情報を明確にする。
保存場所、ファイル名の命名規則を明確にする。
各データの列はタブで区切る。

◎分析データファイルの設定
分析データファイルを探す機能を付ける。
検索機能、または命名ルールを工夫しユーザが見つけやすいものにする。

◎相関ルールの選択方法について
特定の空間、時間に関しては選択できない。
例えばアメリカの北側、1990年以降のみ等の指定は出来ない。

◎KMLの相関ルールの表示方法
大気汚染以外のデータにも対応できるようにするため、日本語の固定メッセージは避ける。
ルールの評価をユーザが判断できるように、具体的な数値を書く。
抽出された相関ルールをそのまま表示する。
ユーザの利便性のため、汚染物質毎に表示オプションを付ける。
確信度に応じて、プレイスマークのカラーや大きさを変えて視覚的に理解しやすくする。

◎今後の予定
11月中旬には詳細仕様を決める。

◎今後書くドキュメント
設計クラス図、設計シーケンス図、内部設計書

●アクション項目：
◎スケジュールの担当と必要な日数を明確にする。
◎今回いただいたご意見に基づき、各ドキュメントを修正する。
開発構想書、画面定義書、画面遷移を修正する。

●決定事項：
特になし

●次回の予定
次回は11月9日(水) 20時から

研究開発プロジェクト議事録

●日付：2011/11/09　20:30～21:30
●場所：インターネット(Skypeを利用したミーティング)

●作成者：沈

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、永田、田中、沈

●フォローアップ案件：

◎前回の修正した部分の検討

◎画面遷移内部処理
内部設計書に書く。

◎リスクの対策
プロジェクトのリスクを識別し、分析し、リスクに対応するための対策は必要です。

◎データの形式
内部設計書に書く。

◎開発構成書の実行時間の測定
相関ルールについて、どなん時間の測定は要らない。
しかし、報告書に事例を持ち、詳しく説明する。

◎ヒストグラムについて
ヒストグラムの表示は５分間ぐらいかかる。

●アクション項目：
◎今回いただいたご意見に基づき、各ドキュメントを修正する。
スケジュール、開発構想書、画面定義書、画面遷移を修正する。

●決定事項：
次回の全体会議は11月26日（水）19:00
研究開発プロジェクト議事録

●日付：2011/11/16　19:00～20:11
●場所：インターネット(Skypeを利用したミーティング)

●作成者：永田佑輔

●参加者：
土田様（日立）、天笠先生、永田、沈、田中

●会議内容：
本会議では、開発構想書と画面遷移図、シーケンス図について、意見をいただいた。
いただいた意見を以下に示す。

◎システム規模の見積り
サンプルプログラムを元に、システム規模を見積もる

◎開発構想書
【1.2.1.開発の目的】
〇「26日間かかっている」という表現から、「26日間かかることが報告されている」に変える。また、Hadoopを入れることで、処理時間がどう変わったのかも書く
〇大規模なデータ分析処理の事例として、Yahooの事例を書いているが、データを時空間に分割して、分析をしているのか？
データを時空間に分割し、分析を行っていないのであれば、時空間で分割し、分析を行っている事例のほうが良い
〇上から三つ目のパラグラフで、主語と述語の組み合わせがおかしい
〇相関ルール分析について、書けてない
相関ルール分析で、有用なルール抽出できることを書く
なぜ、相関ルール分析を行うのか。分析ができると、何が嬉しいのか。
〇文書構成としては以下の流れで書くと良い。なお、以下の項目は一般論で書く。
1.分析のニーズ
2.相関ルール分析
3.ログデータを相関ルール分析を行うニーズ
4.分析にかかる時間
5.分析時間の高速化(Hadoopで並列処理をする)

【2.想定する利用者】
〇大量のログデータを持っていて、時間軸と空間軸に分割し、相関ルールを分析を行い、結果を視覚化する利用者を想定している
〇利用者が持っているデータ、行う分析方法を書く

【3.開発するシステム】
〇基本用件が抜けている
基本用件は利用者がシステムをどう使うのかを書く
誰のために、どう使うのか。利便があるのか等。

【5.ドキュメント要求】
〇ドキュメントを書かない場合は、その理由も書く
〇内部設計書はチーム内でモジュールのフロチャートとクラス図とシーケンス図をチーム内で共有するために書く
〇ドキュメント要求には本当に必要なものだけを書く
ドキュメントの意味を書いておく

◎画面遷移図
以下の点を修正し、確認した。なお、修正箇所は以下の通りである。
〇データ加工システム画面に「ファイルリスト表示」ボタンを付加
〇データ加工システム画面に「出力ファイル名」の入力テキストボックスを付加
前回の会議まで、出力する中間データのファイル名をシステムが自動で決めるようにしていたが、使用性を考え、利用者が入力できるようにした。

◎シーケンス図
画面周りのやり取りが必要がない。例えば、スライドバーを動かしたら、パラメータの再設定がいる。

●アクション項目：
◎スケジュールの定量的な見積もり
◎リスクの識別とその対応
◎ドキュメントの指摘事項を修正

●決定事項：
特になし

●次回の予定
次回は11月24日(木) 19時から

研究開発プロジェクト議事録

●日付：2011/12/01　19:00～20:30
●場所：インターネット(Skypeを利用したミーティング)

●作成者：田中節子

●参加者：
土田様（日立）、天笠先生、永田、沈、田中

●会議内容：

◎特定課題研究報告書の主題と副題の決定
〇主題について
Hadoopを用いた大規模ログデータに対する相関ルールマイニングシステムの開発
→学生側が主題の意味を説明できるように。

・Hadoopがメインなので、タイトルが明確になるようにRはあえて書かない。
・大規模ログデータは時系列データを含んでいるため、一般的な名称を使用した。
・相関に関する誤解を招かないため、相関分析システム→相関ルールマイニングシステムに変更した。

〇副題について
永田：プロジェクト管理とデータ加工システムの開発
田中：ソリューション企画とデータ分析システムの開発
沈：ユーザブルなUI開発

自分が一番力を入れたところを書く。
12月1日中にメールで先生方に送信する。
この研究開発プロジェクトを通じて得たことを特定研究課題報告書に書く。
トラブル→あれこれやった→解決した

◎見積もりについて
見積もりの行数がどういう意味を持つのか、明確にする。
画面とバッチでは1行の意味が違う。
JSPで何行、JAVAで何行、と明確にする。

◎リスクについて
プロジェクトに対する評価を報告書に入れた方が良い。
どこは良いが、どこはもっとやれば良かった。
どの機能はあれば良かった。

◎開発構想書について
多様なデータ形式に対応するために、プラグインを用いて入力データの入出力を制御する。
タイトルに関係することは統一する。
プログイン機能は、入力データに特化する部分をさまざまなデータに対応する機能である。
ソフトウェア構成の図は修正する。JDK、JREは削除する。
ソフトウェア構成にGoogleEarthを追加する。
KMLについての説明を追加する。

●アクション項目：
◎頂いたアドバイスをドキュメントに反映する

●決定事項：
特になし

●次回の予定
次回は12月9日(金) 19時から

研究開発プロジェクト議事録

●日付：2011/11/16　19:00～20:11
●場所：インターネット(Skypeを利用したミーティング)

●作成者：永田佑輔

●参加者：
土田様（日立）、天笠先生、永田、沈、田中

●会議内容：

外部設計書

・2.2.で異常処理も書いたほうが良い。
　エラー処理は明示的に書いたほうがよい。
・プラグインが合わなかった場合はどうなるのか。
・加工も分析も、時間がかかる処理なので、気になった。
・入力画面という名前が不適切。
・google earthの連動はなし。

・ER図を見ると、このシステムは大気汚染データのみを扱うみたいに見える。
・どちらかというと、内部設計書に書く。

・プラグインの仕様はこういうのが必要ですのを書く。

内部設計書
・有効値の範囲はマスタファイルから読み込んで行う
・閾値はプラグインを通さない
・個別のファイル処理はプラグインの内部処理に書く
・SiteIDの

一日なら
　一かいでも超えてる場合と24時間超えてる場合と、過半数を超えてるのか

[20:28:35] nagata: 外部設計書

・2.2.で異常処理も書いたほうが良い。
　エラー処理は明示的に書いたほうがよい。
・プラグインが合わなかった場合はどうなるのか。
・加工も分析も、時間がかかる処理なので、気になった。
・入力画面という名前が不適切。
・google earthの連動はなし。

・ER図を見ると、このシステムは大気汚染データのみを扱うみたいに見える。
・どちらかというと、内部設計書に書く。

・プラグインの仕様はこういうのが必要ですのを書く。

内部設計書
・有効値の範囲はマスタファイルから読み込んで行う
・閾値はプラグインを通さない
・個別のファイル処理はプラグインの内部処理に書く
・SiteIDの

一日なら
　一かいでも超えてる場合と24時間超えてる場合と、過半数を超えてるのか
[20:37:33] tanaka: ・2.2.で異常処理も書いたほうが良い。
　エラー処理は明示的に書いたほうがよい。
・プラグインが合わなかった場合はどうなるのか。
・加工も分析も、時間がかかる処理なので、気になった。
・入力画面という名前が不適切。
・google earthの連動はなし。

・ER図を見ると、このシステムは大気汚染データのみを扱うみたいに見える。
・どちらかというと、内部設計書に書く。

・プラグインの仕様について書く。

内部設計書
・有効値の範囲はマスタファイルから読み込んで行う
・閾値はプラグインを通さない
・個別のファイル処理はプラグインの内部処理に書く
・プラグインの具体例としてアメリカのデータで説明する。

[20:38:40] tanaka: プラグインの必須項目を書く。例SiteID
[20:39:41] nagata: 属性の順番は関係ない
[20:50:37] nagata: DataProcessSystem
[20:50:52] nagata: DataAnalyzeSystem
[20:51:23] nagata: DataProcess
[20:51:32] nagata: DataAnalyze

●アクション項目：
◎スケジュールの定量的な見積もり
◎リスクの識別とその対応
◎ドキュメントの指摘事項を修正

●決定事項：
特になし

●次回の予定
次回は11月24日(木) 19時から

研究開発プロジェクト議事録

●日付：2011/12/14　20:20～21:30
●場所：インターネット(Skypeを利用したミーティング)

●作成者：永田佑輔

●参加者：
土田様（日立）、天笠先生、永田、沈、田中

●会議内容：

20:20～
参加者：
土田さん、西澤さん、天笠先生、

開発ドキュメントについては
開発メインに行ったので、完全に修正していない。

以前のドキュメントで指摘事項があれば、
おっしゃってください

データ集計とKMLファイルを担当している
今は分析システムのデータ集計

12月22日に報告書の第一版を書かないといけない

12月18日までに、

12月18日までに個人が担当している部分を終わらせる

山戸先生に言われたことを日立製作所の人と天笠先生に相談すること

サブシステムごとに結果が取れるのはいつごろ？
全体的なシステムの結果が取れるのはいつごろ？

ここまでで、完成のイメージがずれていそう

開発のイメージは？

天笠先生のレビューを受けて、主査のレビューを受けて、
最後に副査に見せる

色々な先生がいるので、まず、天笠先生に相談する

報告書は読み手がいるので、目次レベルでレビューする
何を書くかを決め手から、スケジュールを決めて、書いていく

報告書の目次は西澤さんと土田さん、天笠先生に
見せてから山戸先生にお見せする

早めに山戸先生に見てもらって、連絡する。
メールを会議後すぐに書く。

前倒したことでの懸念
　・検討時間が無くなる
　・体への影響が心配

データ分析システムで、
オブジェクトの結果をファイルに書き出せばいい
これは技術調査でやること

技術調査　→　設計　→

データの集計作業でHadoopを使う
分析はRで行う
Rの入力はデータベースから読み込むとかインタフェースはある

課題をチーム共有し、調べても難しい場合は、日立の人に送ってもいい
課題リストを作り、どの課題が解決できたのか共有化したい

ファイル渡しで起こる問題はなんですかね

HDFSに書くと重いので、ローカルディレクトリに書いたほうよい

評価の仕方としては、比較対象のページを作って、比較するの？

html5と新しい技術を使ったページの操作性の比較は？

操作性の評価とは、何人かに使ってもらって、長い間してもらう

この研究だと、何か有益なマイニングができたのかが評価基準

こういうことが求められていて、このユーザインタフェースではそれを満たしている

何人かのログをとることはできない。

新しい技術を使いましたは評価できない
新しい技術を使った結果、どうなったのかがいえないとだめ

開発者側からのメリットより、利用者側からのメリットを述べたほうがよい

時間：
　12月20日(火) 20時から

http://www.ne.jp/asahi/hishidama/home/tech/apache/hadoop/tutorial.html#h_debug

●アクション項目：
　◎スケジュールの定量的な見積もり
　◎リスクの識別とその対応
　◎ドキュメントの指摘事項を修正

●決定事項：
　特になし

●次回の予定
　次回は11月24日(木) 19時から

研究開発プロジェクト議事録

●日付：2011/12/20　20:15～21:30
●場所：インターネット(Skypeを利用したミーティング)

●作成者：田中節子

●参加者：
　土田様（日立）、西澤様（日立）、天笠先生、沈、田中

●会議内容：

　◎外部設計書に、ユーザに必要な情報が書かれていない
　　今後、外部設計書に含めること
　　〇プラグインの仕様
　　　・ユーザに用意してもらう内容を明確に書く
　　〇前提条件
　　〇相関ルールの仕様
　　　・3種類のアソシエーション分析から分かることを書く

　◎システムの評価方法についてドキュメント化する
　　〇どのような方法で評価するのか？
　　　・相関ルールの抽出結果に基づいて判断するのか？
　　　・スケーラビリティに基づいて判断するのか？

　◎誰がどの部分の開発担当なのかドキュメント化する

　◎テスト実施報告書を書く

・単体テスト報告書
・結合テスト報告書

◎プログラムは一度に全部つなげるのではなくて、
少しずつ重要な部分からつなげた方が良い

◎今後の課題、今回の開発でやり残した点は、特定課題研究報告書に各自書く。

●アクション項目：
◎外部設計書の修正
◎評価基準をドキュメント化する
◎役割分担をドキュメント化する
◎テスト報告書を作成する

●決定事項：
特になし

●次回の予定
次回は12月27日(火) 19時から
研究開発プロジェクト議事録

●日付：2011/12/27　19:30～20:30
●場所：インターネット(Skypeを利用したミーティング)

●作成者：沈

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、田中、沈

●会議内容：
◎スケジュールが遅れる
対策；１、早めに実装を終わる。
２、他人の同じ機能のソースを参考する。

◎実装
システムの構成図を確認。
田中と沈は２８日までに完成。
永田は1月1日までに完成。

◎単体テスト
１月2日に開始

●アクション項目：
◎３０日までに評価書を作成する

●決定事項：
特になし

●次回の予定
次回は1月4日(火) 17時から

研究開発プロジェクト議事録

●日付：2012/01/04　17:39?19:04
●場所：インターネット(Skypeを利用したミーティング)

●作成者：永田

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、永田、沈

●会議内容：

【システム評価書】
・データセットは入手元とレコード数、データ容量を書く
また、各属性の意味と属性にどういうものが入るのかを書く
・どれが固定パラメータか変動パラメータかわからない
・評価項目の中で、行わない項目の理由を書くより、行う項目の理由を書くほうがよい
行わない項目の表現は「以下に行わない項目を示す」程度で良い
・実機の環境設定に時間がかかるとはかかないほうがよい
・優先度については、優先度の基準を記述する

【進捗の確認】
・スケジュールを項目分けして、状態を記述する
状態例：進行中、完了
また、各項目の進捗率を記述する
・プロジェクトの進め方を議論して確認しないといけない
・遅れるのアクションを書かないといけない
・今週の作業内容と次週の作業内容を記述する

●アクション項目：
・現在の状況を反映したスケジュールを関係者に送る

●決定事項：
・作業内容と課題を、適宜、報告する
報告頻度は開発メンバで決める
・統合テストの中日にメールを送る

●次回の予定
次回は1月11日(水) 19時から

研究開発プロジェクト議事録

●日付：2012/01/11　19:00-21:30
●場所：インターネット(Skypeを利用したミーティング)

●作成者：沈

●参加者：
土田様（日立）、西澤様（日立）、天笠先生、永田、沈

●会議内容：

【システム評価書】

評価書
項目
◎沈の相関ルール分析の結果、永田さんと田中さんの結果と比較しない。

◎相関ルールの結果について、条件部の集合はない、結論部があるの原因は何？
調査するは必要

◎粒度は集合のサイズを変る

◎「Google Earth」物質で分けるの仕様を実装するか？
時間はないので、やめます。

◎hadoopコマンドで、Rが実行できない
天笠先生と相談する。

報告書
◎データ加工処理に時間がかかる
VirtualBOXじゃない、３人の独立の機器を集めて、試す。

◎システム２次利用者は、KMLファイルの利用できる。

●備忘：
◎最終発表は３０日13:45開始

●次回の予定
次回は1月19日(木) 19時から

研究開発プロジェクト議事録

●日付：2012/01/19　19:30-20:30
●場所：インターネット(Skypeを利用したミーティング)

●作成者：沈

●参加者：
土田様（日立）、天笠先生、田中、沈

●会議内容：

「システム画面と加工部、分析部の連携」
日曜日まで完成する

「データ加工処理に時間がかかる」
VirtualBOXじゃない、３人の独立の機器を集めて、試す。

「環境構築」
日曜日まで完成する

●アクション項目：
発表練習は1月26日(木)　19時

●次回の予定
次回は1月26日(木) 19時から