Development of a Data visualization and analyze system using Tag Cloud

Abstract

This paper describes about a database Visual Analytics system that uses a Tag Cloud on the basis of a degree of semantic similarity and a coefficient of determination. The system makes data discovery easy by deciding tag size according the degree of semantic similarity. The system also supports data analysis by assigning a color to the tag according the coefficient of determination.

Recently, the amount of data is increasing. And, opportunity to get new knowledge by utilizing data is increasing. So, data utilization is getting important. Data discovery is first step of data utilization. Several techniques have been made on data discovery, and one of the techniques is semantic similarity search technique. Semantic similarity search technique make data discovery easy. But these search technique cannot detecting data perfectly. So, displaying the data clearly which is semantic similar with a search keyword is efficient for detecting data. Displaying the data clearly is important. Data analysis is second step of data utilization. Single regression analysis is a basic part of the data analysis. This reveals the relationship of two data. But this relation is not always true. In order to reveal appropriateness of single regression analysis, coefficient of determination is used. So emphasis high coefficient of determination 's data makes data analysis efficient.

This research conducts a comparative experiment with a spreadsheet program which has many function of data utilization to verify the effectiveness about data utilization. The content of comparative experiment is to do hypothesis testing from one database. Also, an operability of the system is tested by NEM (Novice Expert ratio Method). These tests suggest that the system has high operability and make data utilization easy.

学　士　論　文

題　目　Tag Cloudを用いたデータ可視化分析システムの開発

指導教官　小山田 耕二 教授

京都大学工学部 電気電子工学科 電気工学専攻

氏 名　今井 晨介

平成27年2月13日

目 次

第 1 章　序論 **1**

第 2 章　関連研究 **3**

2.1　検索結果の可視化システム 3

2.2　Tag Cloud を用いたデータベースの可視化 3

2.3　Tag Cloud に意味的類似度を用いた可視化 4

第 3 章　提案システム **6**

3.1　要件定義 6

3.1.1　使用事例 6

3.1.2　ヒアリング調査 6

3.1.3　システム要件 7

3.2　システム設計 7

3.2.1　システム概要 7

3.2.2　機能詳細 9

3.2.3　処理手順 10

第 4 章　実験と考察 **12**

4.1　検証実験および結果 12

4.1.1　NEM(操作性能測定) 12

4.1.2　対比実験 14

4.1.3　ユーザーアンケート 15

4.2　考察 15

4.2.1　NEM(操作性能測定) 15

4.2.2　対比実験 16

4.2.3　ユーザーアンケート 16

4.3　課題・改善案 17

第 5 章　結論 **19**

謝 辞 **21**

参 考 文 献 **22**

付 録 A 付録 A 基礎技術 (WORDNET) **24**

第1章 序論

我が国では、科学技術政策と科学技術に関連するイノベーションのための政策を推進する「科学技術イノベーション政策の一体的展開」を掲げている[1)]。科学技術と社会との関係が深化する中、科学技術イノベーション政策を「社会及び公共のための政策」の一環として、国民の幅広い参画を得つつ、理解と信頼を得ながら進めていくためには、科学的方法に基づき政策の企画立案及び推進等を行い、政策形成プロセスをより合理的なものにするとともに、国民に対してより一層の説明責任を果たしていくことが必要となる。

この科学的方法を利活用する際 , 重要な役割を果たすものが 仮説検証法である[2)]. 仮説検証法では、問題に対して、仮説を構築し、データを用いて検証が可能となるように因果関係を定式化（概念操作化）する。この定式化された因果関係として表現された仮説をデータを用いて検証することで科学的方法に基づいた政策策定が可能になる。この場合、因果関係を構成する原因と結果に対応するデータを観察や調査などを通して収集することになるが、政府統計データ等で既に収集されている場合には、データに付与されたタグを手掛かりに検索することとなる。ここで、タグとは、データベースの列名をあらわす。

近年、政府統計データなどオープンデータの利活用できる環境が整備されつつあるが、仮説検証という観点ではまだまだ使いやすいとは言えない。仮説を構成する原因と結果に対して、概念操作化を行ったとしても, それらに対応するするタグを効率よく見つけることは困難である。たとえば、都市化が進むと犯罪率が高まるという仮説を検証するうえで、都道府県別の犯罪率データを調べるために、「犯罪率」というキーワードで政府統計データベースを検索しても対応するタグは見つからない。「犯罪率」と等価なデータは、「刑法認知件数」というタグで登録されているからである。仮説検証を効果的に進めるためには、オープンデータシステムでは、登録されているデータのタグが分かりやすく俯瞰表示されるのがよい。大量のタグが同じように表示されていても気づきにつながらない可能性が高いので、検索したいデータのタグの異音同義語に対する意味類似度を属性にして表示される仕組みが望ましい。

上述のように, データの関係性に着目してその結果を利用者にわかりやすく提

示することは, 仮説検証の大きな手助けになる. 本論文では, その目的のために, Tag Cloud と呼ばれる可視化手法に着目した. Tag Cloud とは文書や Web Site の可視化手法で, 文書の単語に重要度をつけ, それを基に Tag Cloud 内のタグのフォントサイズやタグの色などを変えた複数の語を並べて表示する手法である. Tag Cloud は, フォントサイズやタグの色の属性を持たせ, それらを適切に設定することで単語の重要度に応じた効果的な表現をすることが可能である.

本論文では, 47 都道府県別の統計データを対象とする, 検索キーワードとの類似度および選択データとの決定係数に基づき Tag Cloud 可視化技術を用いた可視化分析システムを提案する. このシステムでは Tag Cloud を用いてタグを表示する。タグのフォントサイズを検索キーワードとの意味的類似度に基づき表示することで、検索したいデータのタグを発見し易くした。さらに, 二つのタグを選択することで、タグの持つデータを散布図, 回帰直線で表示する。仮説検証では、仮説構築からデータ収集、分析、検証と複数の手順を要するが、そのうちのデータ収集と分析を同時に行うことで、仮説検証の効率化を図った. また、タグの色については決定係数に応じて決定する。回帰直線の蓋然性を示す決定係数をタグの属性にして表示することで、関連の高いデータを探しやすくし、データ関係の分析を通して、関連が高い原因 (共通原因や比例、因果関係など) を新たに発見につながる知的基盤への機能拡張を目指した。これらにより提案システムでは仮説検証の効率化を目指した.

本システムの有効性を検証するために, 本論文では比較実験, 操作性能実験を行った. 操作性能実験により提案システムの操作性を検証し, 回帰分析などデータ分析についての機能を多く持つ表計算ソフトウェアの Excel と比較実験を行うことで, 仮説検証での有効性を検証した. また利用者から実験後にシステムについてフィードバックをもらうことで提案システムの不足している点などについての意見をもらった.

本論文の構成は以下の通りである. 第 1 章は序論である. 第 2 章では関連研究についてまとめる. 第 3 章では提案システムについて述べる. 第 4 章では, 適用例として利用者が提案システムを利用した実験結果を述べ, 考察を行う. 第 5 章では結論および今後の展望について述べる.

第2章 関連研究

本章では, 本論文の関連研究をまとめる. 2.1 節では利用者の検索に基づきデータベースを可視化, 分析するシステムについて, 2.2 節では Tag Cloud を用いたデータベースの可視化について, 2.3 節では Tag Cloud に類似度を用いた可視化について説明する.

2.1 検索結果の可視化システム

データベースを対象とた検索結果の可視化技術はこれまでにもいくつか提案されている. Clarkson ら [3] は文書データベースの検索システム ResultMaps を提案した. この手法では, 階層的なメタデータを持つデータベースの検索結果表示に対してリスト表示に加え, 図 2.1 のように Squarified Treemaps [4] と呼ばれる方法を用いてツリー階層を表現することでデータベースの特徴を分かりやすく表示した. これらの手法では, ツリー構造を Treempas 技術を使うことで, 検索結果を限られた空間にデータを詰め込むことを実現しているが, 提案システムで扱うデータは階層構造を持たないのでといった Treempas 技術を生かすことが出来ず, また,Treempa 表示では TagCloud に比べ一つ一つのデータに対して多く面積を取るのでより多くのデータを表示することには適していない.

2.2 Tag Cloud を用いたデータベースの可視化

Tag Cloud とは, 単語集合に対して, 重要度に応じて単語（タグ）の大きさや色を調整し, 平面上の任意の閉領域内に単語を配置する技術である. Lohmann ら [2] は Tag Cloud による可視化はリスト表示に比べ, 重要度の高いタグの発見を早める傾向があると評価している.

Tag Cloud 内のタグのフォントサイズを変更する例として Wordle [6] が挙げられる. Wordle は単語の文書内で使われた回数 (出現頻度) でフォントサイズに重み付けをし, 中央からランダムに配置し, 文字の隙間や, 空洞部分など重なりさえしな

ければ配置できるのでタグを中心付近に集約して表示する (図 2.2). Wordle を複数の文書に用いることで文章の特徴を明らかにした.

また, Tag Cloud でフォントサイズ, タグの色を変更する例として Chen ら [5] の映画推薦システムが挙げられる. Chen のシステムでは, 英語の説明文を用いて Tag Cloud を生成する. まず利用者が興味のある単語, 興味のない単語を選択する, その後に映画の説明文内でユーザの興味のある単語の出現回数と興味のない単語の出現回数を測定しそれを基にユーザが興味を持つと推測された映画を推薦する (図 2.3). これによって既存のシステムより満足度が高いシステムを提案した.

図 2.2, 2.3 のように Tag Cloud の大部分は Web Site, 文書での単語出現頻度など, データの中身に基づき Tag Cloud を表示しているが, これでは利用者の要求を考慮に入れておらず, 利用者の求める情報を手に入れにくい場合がある [7]. 本論文ではユーザの検索したキーワードとの類似度の高さを基にフォントサイズを, 及びユーザの選択したデータと単回帰分析を行った際の決定係数の高さに基づきフォントサイズとタグの色を決め, Tag Cloud で表示している. これによって各ユーザに適した Tag Cloud 表示を行っているその点で本節でのシステムとは異なる.

2.3 Tag Cloud に意味的類似度を用いた可視化

本節では, 単語間の意味的類似度を基にしてフォントサイズやタグの色を決定する Tag Cloud 可視化技術について述べる. 意味的類似度とは単語間の意味的な類似性を定量化したものであり, 意味的類似度の計算方法は複数存在する. Tessem ら [8] は現在地について利用者が興味を持つ情報を可視化するために Android 用アプリケーションを開発した. Tessem らはアプリケーションを使用する情報端末から使用時の位置情報を取得し, 現在地の情報を DBpedia [10] から取得し, その中の単語を, 文中での出現頻度を基にフォントサイズを決定し, Tag Cloud 表示する. また次にタグを選択すると選択されたタグと全てのタグの意味類似度を算出し, 意味類似度と出現頻度の積でフォントサイズを決定する. このシステムでは, 文中での単語出現頻度だけでは地理的語が多く表示されるが, 利用者に興味のあるタグを選択させることで利用者により価値のある Tag Cloud を生成することが可能になった (図 2.4).

また, Wang, ら [9] は Amazon 等の商品に書かれるレビュー記事の分析に Tag Cloud を用いた. 記事内に存在する単語同士の意味的類似度を測定し, それを力

学モデルのバネ係数に割り当てた．これにより, 似た意味の言葉との距離は近くに集まり, レビュー記事の特徴を認識しやすくなった. フォントサイズはレビュー文内での出現頻度で, タグの色はクラスター別で決定する (図 2.5).

上記のように Tag Cloud に意味的類似度を用いることで利用者にとってより価値のある情報を表示できる．しかし, 提案システムでは利用者の入力キーワードとの類似度を利用している．検索を利用することで, 限られた選択肢以外の要求を行うことが可能になった．また Tag Cloud で可視化する対象が多くの手法では文書であるが, 提案手法では単語を扱っている点も異なる. 文書では単語の文書内での出現頻度を指標にすることが多いが, 提案手法では単語を扱うので出現頻度を扱うことはなく, 代わりに検索キーワードと類似度の高さ, 及び選択データと単回帰分析を行った際の決定係数の高さを用いる.

第3章 提案システム

3.1 要件定義

3.1.1 使用事例

本項では, 仮説検証補助システムについて考えうる使用事例を通して, 研究目的である仮説検証の効率化に必要である要件を分析した.

まずは, 2 つの現象の関係性についての仮説を立てる. その後, インターネット上で公表されているデータベースから所望のデータを検索する. 2 つの現象名と一致するデータが取得できない場合は, 2 つの現象名をデータから検証が可能となりやすい名前に構成しなおす必要がある. 2 つの現象名と一致するデータを手に入れた場合, データを用いて, 説明変数と目的変数の式 (回帰式) を生成する. 生成した後, 仮説についての蓋然性の高さを確かめるための指標である決定係数を参考にする. この値が低い場合, 仮説の蓋然性が低いということで仮説を構成しなおす必要がある. 以上の流れの中で, 仮説検証の補助に必要な機能について述べる. 仮説の変数概念が素早く見つけることができるように, 2 つの現象名と一致するデータ, もしくは代わりになれる可能性が高いデータを分かりやすく表示する必要がある. このためには, 2 つの現象名とどれだけ似ているかを分かりやすく表示しつつデータを表示する必要がある. また, 回帰式・決定係数を参考にし仮説検証を行うため, 回帰式・決定係数を分かりやすく表示しておく必要がある. このために少ない操作ステップで, 回帰式, 決定係数を表示する必要がある.

3.1.2 ヒアリング調査

データ活用についての機能を多く持つ Microsoft Excel(Excel) を用いて仮説検証を行た際に, 利用者へヒアリングを行った. 以下にその結果をまとめた.

- 機能が多すぎて自分の求める機能を見つけづらかった
- あまり Excel を使ったことのない利用者にとって散布図のデータ変更, 回帰分析などが複雑だった

これより, 必要最低限の機能数かつ, 少ない操作ステップで扱える機能が必要だと分かった.

3.1.3 システム要件

ここまでの使用事例, 利用者へのヒアリングをふまえて，提案システムへの要件を以下のように想定した.

- 検索キーワードとの意味的類似度が分かるようなデータ一覧表示
- 選択データとの決定係数が分かるようなデータ一覧表示
- 回帰式, 決定係数の表示
- 少ない操作ステップ

3.2 システム設計

3.2.1 システム概要

前節でまとめたシステムの要求要件をもとに，本研究で提案するシステムの概要を説明する．提案するシステムは以下の機能から構成されている．各機能の詳細は後述する.

- 利用者の検索キーワードに対するデータの意味的類似度算出
- 算出された類似度に応じフォントサイズを決定し, Tag Cloud で表示
- 利用者のデータ選択に対する他データの決定係数取得
- 取得した決定係数に応じタグの色を決定し, Tag Cloud で表示
- 利用者の選択した二つのデータの散布図, 回帰直線表示
- 少ない操作ステップ

図 3.1 はこれらの機能の関係を示している．以下では図 3.1 の流れに沿って提案システム利用の流れを概説する.

まずシステムを起動すると図 3.2 のような画面が出力される．利用者は求めているデータ名についてのキーワードを検索欄に入力する．検索を始めると, 検索キーワードとタグの意味類似度を算出し, その値に基づき Tag Cloud のフォントサイズを決定する. 図 3.3 のように意味類似度が高いタグのフォントサイズを大きく表示される (図 3.3 では「犯罪」で検索を行った). 次に表示された中からタグを一つ (散布図 X 軸用) 選択する. 選択をすると図 3.4 のように全てのタグは選択されたタグの持つデータに対する決定係数を取得し, その値に基づき Tag Cloud のタグの色を決定する．決定係数が高いタグは青色, 低いタグは赤色で表示される (図 3.4 では「刑法犯」を選択し, 各タグは「刑法犯」との決定係数に基づき色付けされた)．それと同時に選択されたタグが持つ 47 都道府県毎のデータを日本地図にマッピングする (図 3.4 では「刑法犯」についてのデータが日本地図にマッピングされた). 更に一つデータ (散布図の Y 軸用) を選ぶと図 3.5 のように, 新たに一つ日本地図にマッピングし, 選択した二つのデータを散布図に回帰直線, 決定係数を加えた図を描写する (図 3.5 では「刑法犯」と「窃盗」についての散布図, 回帰直線を表した)．X 軸, Y 軸どちらかのデータを変更する場合は軸変更ボタンを押し変更したい軸を選択した後 (図 3.6), 変更したいタグを選択することでデータを入れ替えることが出来る．また, この際 Tag Cloud のタグ色は変更されない軸のデータとの決定係数に応じて決定され, 表示される. X 軸, Y 軸共にタグが選択されている時点で新しいタグを選択することで, タグが変更された軸についての日本地図は上書きされる.

システム開発

本システムは主に JavaScript を利用して開発した．また類似度を計算する際に日本語 WordNet, 英語 WordNet, また Python, PHP を利用した. WordNet とは単語の上位 / 下位関係, 部分 / 全体関係, 同義関係, 類義関係などによって単語を概念 (synset) に分類し体系づけた概念辞書 [11] で, 日本語 WordNet は日本語の, 英語 WordNet は英語の概念辞書である．英語 WordNet には synset 間の最短距離を測定する API があり, これを基に提案システムでは類似度を算出する．そこで提案システムでは, 検索キーワードの synset の ID を取得するために日本語 WordNet を, synset 間の最短距離を測定するために英語 WordNet を利用した.

使用データベース

今回の実験では政府統計の総合窓口 (e-Stat) 内に登録されているデータからデータベースを作成した. e-Stat とは各府省等が持つ統計データを集約させた政府統計ポータルサイトである. 今回使用したデータは, 種類が多かった 2010 年の 47 都道府県データ 90 個を選択した. e-Stat から取得したファイルのフォーマットは Microsoft Excel 用のものであるが, システムで利用できるように図 3.7 のような json ファイルに変更し, システムに使用した.

ここで扱われるデータには図 3.8 のように各自, データ名「table_name」, データのタグ「tags」, 47 都道府県のデータ値の集合「data」, 各データとの決定係数「correlation」が含まれており, データの集合をデータリストと呼ぶ. またタグは一つのデータ毎に一つ, それぞれ異なるタグを持っており, これはシステム設計者が作成した.

3.2.2 機能詳細

WordNet を用いた類似度算出

利用者の入力したキーワードと各タグとの類似度を WordNet を用いて算出し, 類似度を基にフォントサイズを決定する. 二つの単語 w1,w2 の単語間の関係の略図を図 3.9 に示す. WordNet の持つ階層構造を用いて語と語の最短距離 d を測定し, 類似度を計算する. 類似度は以下のようにして求める.

$$(Similarity) = \frac{1}{1+d} \tag{3.1}$$

単語が複数の synset に属する場合, すべての synset での最短距離を測定し, その中で単語間距離が一番短くなる値を最短距離とする. また, 類似度辞書に存在しないデータとの類似度は 0 とした.

類似度によるフォントサイズ決定

前項での類似度算出を全てのタグに対して行い, 算出された類似度に応じてタグのフォントサイズを決定し, Tag Cloud で表示する. 利用者の入力したキーワードに類似するタグのフォントサイズを大きくすることで利用者の求めるデータの発見を容易にする. フォントサイズの範囲は予め固定し, 類似度の二乗の値を線形

スケールして決定する．類似度を二乗してフォントサイズを決定することで類似度の高いデータがより強調される．

決定係数によるタグの色決定

Tag Cloud 内のタグを一つ選択すると，他のタグが選択タグの持つデータとの決定係数を取得し，それに応じてタグの色を変更させる．データ間の決定係数は予めデータ内に登録されている．また，取得した決定係数に応じてタグの色を決定する．利用者の選択したタグと関係の強さに基づきタグの色を決定することで利用者がデータ分析を行う際にデータ間の関係をつかみやすくする．タグ色の範囲は予め固定し，決定係数を線形スケールして決定し，Tag Cloud で表示する．

選択データについての散布図，回帰直線，決定係数

二つのデータを選択することで，二つのデータの散布図を描写し，同時に回帰直線を描写する．散布図内には二つのデータによる単回帰分析を行った際の決定係数を表示し，散布図の点にマウスオーバーすることで点の持つ X 軸，Y 軸の値を表示する機能を付けた．また，回帰直線によって，従属変数 Y が説明変数 X によってどれくらい説明できるのかを定量的に分析することができる．回帰直線は次式で表す．$Y = aX + b$ また係数 a, b は以下のように最小二乗法を用いて定義される．

$$b = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \tag{3.2}$$

$$a = \overline{y} - b\overline{x} \tag{3.3}$$

x_i, y_i は各都道府県の X 軸，Y 軸に対する値 $(1 \leqq i \leqq 47)$, $\overline{x}$, $\overline{y}$ はそれぞれ変数 x_i, y_i の平均値である．

3.2.3 処理手順

システムが立ち上がり，Tag Cloud を生成するまでの処理は JavaScript を用いて行われている．これ以降の処理手順を説明する．

キーワード検索を行ってから Tag Cloud を生成するまでの流れ (図 3.10)

検索キーワードを Web サーバに送信する．このキーワードと使用データベース内のデータ名を PHP 経由で Python に送信する．Python の持つ API を利用し, 日本語 WordNet 内でのキーワードとデータ名の概念 ID を取得する. この概念 ID を用いてキーワードと使用データベースのデータ名の英語 WordNet 内での最短距離を取得し, 最短距離から類似度を算出する. その後, 類似度を Web ブラウザが受信し, JavaScript を用いて類似度からフォントサイズを決定し, Tag Cloud を生成する.

タグ選択からタグの色が変更するまでの流れ (図 3.11)

Tag Cloud 内のタグを選択することで, タグの文字列を Web サーバに送信する．Web サーバ内の使用データベースからタグの文字列を tags に持つ配列を選択する. この配列から決定係数を取得し, この決定係数を Web ブラウザが受信し, JavaScript を用いて決定係数からタグの色を決定し, Tag Cloud 内のタグの色を変更する.

タグ選択から日本地図, 散布図, 回帰直線, 決定係数表示までの流れ (図 3.12)

Tag Cloud 内のタグを選択することで, タグ名を Web サーバに送信する．Web サーバ内の使用データベースからタグの文字列を tags に持つ配列を選択する. この配列から 47 都道府県別のデータ値を取得し, このデータ値を Web ブラウザに送信する．このデータ値を基に JavaScript を用いて日本地図にマッピングする. また, Web ブラウザがデータ値を取得した直後に X 軸, Y 軸がデータ値を持っている場合は, そのデータ値を基に JavaScript を用いて散布図を生成する．同時に JavaScript で値を基に回帰式を計算し, 回帰直線を散布図に生成する. また, 取得済みの決定係数から X 軸, Y 軸のデータについての決定係数を選び, それを散布図内に出力する.

軸変更からタグの色が変更するまでの流れ (図 3.13)

軸変更ボタンを押すことで, データ変更されない方の軸に対する決定係数が, 軸変更ボタンを押す以前で取得済みなので, この値を基に JavaScript を用いて決定係数からタグの色を決定し, Tag Cloud 内のタグの色を変更する.

第4章 実験と考察

4.1 検証実験および結果

本節では, 提案システムの有効性を検証するために行った実験の説明を行う.

4.1.1 NEM(操作性能測定)

本項では, 提案システムの操作性を評価するために行った NEM について説明する. システムの操作性は, 単に作業時間の短縮を可能にするだけでなく, 意図通りに操作できないことでの思考の断絶, ストレスを防ぐためにも重要となる. 今回行った NEM は Novice Expert ratio Method の略称で, システムを使用するときの Expert(熟練者・設計者) と, Novice(初心者・一般利用者) の操作時間を比較することで定量的にシステムの問題点を発見する操作性評価手法である. また, NEM は 2009 年に発行された電子政府ユーザビリティガイドライン [12] で, Web サイトの利用品質 (ユーザビリティ) の定量的な目標値の一つとして取り上げているように, あるシステムのユーザビリティの指標として有効である. NEM では熟練者・設計者と初心者・一般利用者の操作時間を, それぞれ t_e , t_n として, NE 比 R を以下のように定義する.

$$R = \frac{t_n}{t_e} \tag{4.1}$$

NE 比 R は, 通常 1 以上の値となり, R が 4.5 以上となる場合は, 何らかの操作性上の問題が含まれていると判断されている [13]. また R が 4.5 以上となる場合は, 両者の操作モデルが一致しており, 操作性に問題がないと判断できる.

測定方法

本実験では, 被験者として大学生 5 名がシステムを利用した. 被験者には提案システムを用いてタスクを行ってもらい, タスク・タスク内の操作ステップに要した時間を測定し, 著者も同様のタスクを行うことで NE 比を測定した. 提案システムでは主に仮説検証を行うが, 二つにデータの探索や, ある一つデータと単回帰分析

を行った際の決定係数の高いデータの探索など複数の情報検索のルートが存在する. これより本実験では, 以下の二つタスクを実行した.

タスク 1：二つのデータの関係性の測定

利用者が持っている仮説 (例：外国人人口が増えるほど犯罪率が高い) を検証するため, 二つのデータを検索し, 検索結果をもとにデータを二つ選択し, 関係性を検証する. 測定時間について, システムが起動してから二つ目のデータを選択するまでの時間を計測した.

タスク 2：一つのデータと関連の高いデータの探索

利用者があるデータと関係が高いものを調べる (例：中学校の数と関連があるデータを調べる) ため, 一つのデータを検索し選択し, その後に関連度の高いデータを選択する. 本実験ではシステムが起動してから, 選択したデータと単回帰分析を行った際の決定係数が 0.8 以上のデータを選択するまでの時間を計測した.

また, タスク内では以下の三つの操作ステップで要する時間を計測した.

ステップ 1 類似度検索を行う.

システムを起動してから検索欄にクリックするまでの時間

ステップ 2 軸を変更する.

Tag Cloud 内のタグをクリックしてから軸変更ボタンをクリックするまでに要する時間

ステップ 3 軸を変更する.

Tag Cloud 内のタグをクリックしてから軸変更ボタンをクリックするまでに要する時間

これらの操作ステップは次のようにタスクに含まれる.

タスク 1：ステップ 1, ステップ 2, ステップ 1, ステップ 2

タスク 2：ステップ 1, ステップ 2, (ステップ 2…)

ステップ 3 についてはタスク 1 の操作後軸を変更し, X 軸のデータを変更するという操作を行わせ測定した. 今回の操作ステップの時間測定では検索キーワード, 類似度計算中の時間を通信時間の誤差の影響を取り除くために省略した. 実験に

は OS が Windows7 の Let's note S9 CF-S9KYFEDR (Core i5, 520M, 2.4GHz/2 コア) を使用し, 操作時間の計測ソフトウェアとして BB FlashBack Express レコーダを使用した.

実験結果を図 4.1 に表す. NE 比についてタスク 1, タスク 2 では 4.5 を下回った. また, タスク 2 について全利用者が決定係数 0.8 以上のデータを一回で取得することができた. 各操作ステップについて, 操作ステップ 1, 操作ステップ 2 は 4.5 を下回ったが, 操作ステップ 3 は 4.5 を上回った.

4.1.2 対比実験

本項では Excel との比較実験の説明をする. Excel は表計算ソフトであり, 回帰分析を始め, 多くの分析機能を備え持っており, データ分析を行えるソフトの 1 つである. 本実験では提案システム, Excel を用いて仮説検証を行い, 実験後にアンケートを実施し, 仮説検証に要する時間, システムの操作性, データの発見のしやすさを調査した. ここでの仮説検証とは, ある二つの事象の因果関係を仮説として立て (例：都市化が進むほど犯罪率が高い), その仮説が正しいかデータを用いて検証することである.

実験結果は表 1 に記した. システムの使いやすさ, 利用者が求めるデータの発見のしやすさ, 仮説実験に要した時間など全て提案システムが上回る結果となった. また, Excel の機能ごとの使いやすさについては, 回帰分析が 3.90, 散布図生成が 4.52, 回帰直線描画が 4.41, 軸変更が 3.48 となった. これに対し提案システムについて軸変更の使いやすさは 3.67 となった.

測定方法

本実験では, 大学生 102 名に対して対比実験を実施した. 上記のように被験者は仮説を作り, その因果関係の真偽を確かめることで仮説検証を行った. 被験者のうち, 73 名は仮説検証に提案システムを用いて行い, 残り 29 名には Excel を用いて行った. 実験を行う前に提案システムの使い方, Excel の持つ機能の一部 (回帰分析, 散布図, 回帰直線生成) を説明した後に, 提案システム・Excel の持つ機能を自由に使い仮説検証を行った. 仮説検証を行った後, システムに対する使いやすさ, データ発見などについて 5 段階 (1 が悪く 5 が良い) のアンケートを実施した.

4.1.3 ユーザーアンケート

提案システムの評価を行うために，4.1.1 項, 4.1.2 項の実験で提案システムを使った利用者に実験後に，システムについてのアンケート, フィードバックをもらった.

ユーザーアンケート結果は表 2 のようになった. すべての項目が中間値の 3 を上回っていたが, 類似度計算時間や類似度計算語彙数, 軸変更の分かりやすさが他の項目と比べると低かった. また, 日本地図は回帰直線, 決定係数, 散布図と比べ仮説検証に役立たなかった.

4.2 考察

実験結果は以下のようになった.

4.2.1 NEM(操作性能測定)

この実験より, 提案システムの利用者は二つのデータの関係性の測定, データと関連の高いデータの探索のタスクについては, 設計者と操作時間は大差なく, 設計者の意図に近い操作を行うことが観測された. 個々の操作ステップを見ると軸変更のステップの操作性に問題があることが分かった. 原因として, 軸変更についての説明が書いておらず, 軸変更ボタンの存在も分かりにくいことだと考えられる. また, 軸変更のステップについては, 操作ステップごとに表示される結果を見ていて操作することを数秒忘れている被験者が見受けられた. 他にも, 本実験では通信時間, タイピングによる誤差を取り除くため, 検索キーワード入力中, 類似度検索中などの時間を測定時間から省いたが, 類似度検索中などの測定時間外の間に次の操作ステップ (データ探索) に備えマウスの移動を行っている者がおり, ここでも誤差が生じた可能性もある. 各タスク, 操作ステップについて, 時間測定タイミングに誤差が生じた可能性が存在する. この時間を考慮すれば NE 比は変化すると考えられる. また, 利用者の実験中の操作を見ることで以下のような課題を発見した.

- タグクリックの際にタグとタグの境界が分かりにくく, うまくクリックできていなかった.
- タグの配色が悪く決定係数が高いデータを選択するのに少し時間がかかった.

4.2.2 対比実験

アンケート結果について評価の差が有意差であることを証明するためにノンパラメトリックな統計学的検定の一つである Mann-Whitney [14] の U 検定を行った. 本実験について「システム全体の個々のアンケート項目の評価について Excel と提案システムは同じである.」という帰無仮説を設定し, 検定を行った所, 表 1 のようにシステムの使いやすさについての P 値は 0.013, 利用者が求めるデータの発見のしやすさについては 0.012 と棄却域の 5% を下回ることから有意差があることが証明された. これより提案システムは Excel より仮説検証の際に使いやすく, データの発見にも優れていることがわかった. また図 4.2 のように, 本実験を受ける以前から Excel で回帰分析, 散布図表示をしたことがある学生が 7 割で, 提案システムは被験者全員が初めて利用した. これより Excel の使用経験のある利用者にとっても提案システムのほうが使いやすいことが分かった. 機能別での操作性については有意差を得ることが出来なかったのでシステム全体の操作性の差がでた要因は分からなかった. これは操作ステップ数が Excel より提案システムは少なく, そこが使いやすさに影響を与えたと考えられる. 仮説検証に対する提案システムの有用性についても, 提案システムを用いた被験者は Excel と比べ仮説検証に要した時間は短く, 散布図生成は多かった. これも Mann-Whitney の U 検定より有意差があると証明された. これは提案システムを用いることで短い時間でデータ間関係を調べることができ, それが仮説検証の効率化につながっていると考えられる.

4.2.3 ユーザーアンケート

類似度, 決定係数を Tag Cloud で表示することは適切だったが, 決定係数に基づいた文字表示については配色への評価が他に比べ低かった. 原因として決定係数によるタグ色の変化が少なく, 決定係数の差が少ないタグ同士だとどちらが高い決定係数を持つのか分かりにくくなる事が原因だと考えられる. また, 類似度検索について度検索時間, 語彙数も課題が見つかった. データ分析機能について, 日本地図は他の機能に比べ役に立たなかったという評価を得た. これについては二つのデータを比較することが目的にも関わらず, 一つのデータについて表示する機能はあまり必要性が感じられなかった. データの軸変更については, タグをクリックすることで軸変更が行われることが分かりづらかった. これについては軸

変更についての説明がシステムに記載されていなかったことに問題があるとえられる.

またシステムの使いにくさについて得られたフィードバックを以下に示す.

- 類似度検索中が利用者に伝わりにくい
- システムの立ち上り時間, 類似度検索時間が長かった
- 配色が微妙で分かりにくい.
- タグの選択がしづらい
- 選択したデータがどれなのかタグクラウド内で分かるようにしてほしい
- 日本地図, 散布図に見切れが生じた
- 外れ値を取り除く機能がほしかった
- 潜在変数を考慮するため, データ同士を乗除出来る機能がほしかった
- 特定の都道府県が散布図のどのあたりに位置しているのかを探すのが困難
- 回帰直線の方程式や次数上げもオプションがあってもよい
- PC によって動作しないものがあった

4.3 課題・改善案

提案システムの課題, 改善点として以下の五点挙げられる.

類似度検索

現在の提案システムでの類似度検索では WORDNET 内の語句しか類似度検索を行えない. この検索方法の検索スピード, 検索語彙数の評価が高くなかったことより改善が求められる. 類似度検索可能な語数を増やすには国立国語研究所発行の分類語彙表など他のシソーラスを用いる. または, 検索エンジンで検索して, 得られたスニペットから語の出現頻度情報を基に確率検索モデル TF-IDF 法など他の方法を用いて類似度を測る方法がある [15]. また類似度計算の速度を上げる方法としてキャッシュすることで計算する時間を省略する方法が考えられる. また類似

度検索中というのが利用者に対して伝わらない可能性があるので検索中はロードアイコンを入れるなどして伝わるようにするべきである.

タグクラウド表示

実験時には, 決定係数を適切に線形スケールしタグの色を決定し Tag Cloud で表示していた. しかしこの決定法では決定係数 0.7 から 1.0 のデータの見分けが難しいように, 決定係数が近いタグの比較がやりづらいという反応があった. この改善案として決定係数を 10 段階に分け, 段階ごとで色を変化させるなどが考えられる. また配色について, 決定係数など一般的な可視化では値が高いものが赤で, 低いものが青色であるのが一般的であるという反応があった. また選択中のタグが分かりにくいということから選択中のタグの色, フォントサイズを変化させるなど今後の改善案として考えられる.

軸の変更軸変更

軸変更ボタンが分かりにくいということで分かりやすい軸変更ボタンの作成が必要となる.

日本地図, 散布図

実験後に日本地図, 散布図に見切れが生じた問題については改善策を行った. また特定の都道府県が散布図のどのあたりに位置しているのかを探すのが困難だったことついても改善した. また, 今後の課題として外れ値対策として, 値の除去をする機能, また潜在変数の影響を取り除けるようにデータ同士を乗除でき, それを X 軸, Y 軸に選択できるような機能も必要である.

PC によって動作しない

古いハードウェア・OS, メモリが少ない場合などに提案システムが動作しないという問題が生じた. しかし詳細な動作環境が調査できていないので今後の課題とする.

第5章 結論

本研究では、オープンデータを用いた仮説検証を支援するために，Tag Cloud を用いたデータタグ検索システムを試作し，その有効性をユーザ評価により検証した．提案システムでは 47 都道府県別の統計データベースを Tag Cloud で表示し，Tag Cloud 内のタグのフォントサイズは検索キーワードとの類似度により決定される．また，データ検索分析をシステム内の簡単な操作ステップで行えるようにすることで科学的思考の中断を防ぐようにした．そしてデータ分析について，データ間の関係を測定することに適している散布図，回帰直線，決定係数を表示し，関係のあるデータの発見をしやすくするために Tag Cloud 内のタグの色は利用者が選択したデータとと単回帰分析を行った際の決定係数の高さに基づき決定される．

このシステムにおけるデータタグの発見容易性を確認するために，本研究では操作性能測定，Excel との比較実験，ユーザーアンケートを行った．操作性能測定により，提案システムを利用して仮説検証を行う際に，初心者でも設計者の動きと近い操作をすることが分かった．比較実験より提案システムでは類似度，決定係数を Tag Cloud で可視化することにより利用者の所望するデータ発見しやすくなり，仮説検証が行いやすくなった．また，ユーザーアンケートから提案システムの改善案を分析した．

また今後の目標としてデータ量，種類の増加が考えられる．提案システムではデータ数は 90 個．また，2010 年度における 47 都道府県別のデータのみを扱ったが，今後は e-Stat 内の全データを使用を検討していきたい．このため大量のデータを提案システムで使用できるよう、専用のデータベースを作成する必要があると考えられる．同時に，データが持つタグ名，決定係数についても改善する必要がある．提案システムでは概念辞書 WORDNET に存在する単語をデータタグに採択したが，データ量，種類が増えることで、タグの採択に負担がかかると予想されるため、今後は負担の少ない採択方法を確立する予定である．決定係数については，事前にデータ同士で単回帰分析を行った際の決定係数を計算していたが，データ量増加につれて計算量が増加し，データ計算の負担は大きくなると予想されるので、システム起動中に計算を行う機能の導入，もしくは決定係数以外の新しい指標を採択するなど対策を検討していきたい．また，データ量が増加に向けて Tag

Cloud の表示方法も検討する必要がある．配置に規則性を持たせなければいけない．タグ同士の類似度によって引力を働かせる, もしくは新しい指標に基づき配置を決定する等データ発見に役立つような配置を考える必要がある．また, e-Stat には 47 都道府県以外にも時系列データ, 市町村別データなどデータ種類の増加に向けた対策として, 動的なグラフや市町村用の地図など、データの種類に対応した地図、グラフの生成機能を追加する必要がある．データ分析については, 現在は単回帰による分析のみだったが、分散共分散構造解析のようなデータ間の因果関係を分析できるような機能を追加する予定である．また, H. Rosling [16] の提案する Gapminder のように時系列の多次元データを可視化分析機能を実装し、さらに 3 次元空間を有効に使った可視化技術を利活用することで、新しい発見につながる知的基盤への機能拡張を検討していきたい.

謝 辞

本研究を進めるにあたり，有益な御指導，御助言を頂きました京都大学高等教育研究開発推進センター小山田耕二教授に深く感謝致します.

本研究を進めるにあたり，豊富なアイデアや活発な議論により，有益な御指導，御助言を頂きました坂本尚久助教に感謝致します.

日頃からの研究の進め方，プログラミング技術や研究生活に必要なさまざまな知識について教えて頂きました久木元伸如先生に心より感謝致します.

本研究を進めるにあたり，プログラミング技術, システム作成等有益なご助言を頂き，多大なる時間を割いて頂きました博士課程 1 回生の尾上洋介氏に心より感謝します.

本論文作成にあたり多くの御助言，御協力を頂きました修士 2 回生の櫛田将史氏, 原大智氏に深く感謝します.

日々の研究室の生活のサポートをしていただき, 本論文の推敲にも多大なる時間を割いて頂きました修士 1 回生の高見円仁氏, 双見恭介氏に心より感謝します.

研究生活のなかで多くの助言と気遣いを頂き，また本論文作成にあたり御指導，御協力を頂いた小山田研究室の皆様に心より感謝の意を表します.

参 考 文 献

1) 科学技術イノベーション政策のための科学推進委員会, 政策のための科学, http://scirex.mext.go.jp/about/index.html (2014)

2) 小山田耕二, 日置尋久, 古賀崇, 持元江津子, 研究ベース学習 (コロナ社, 2011) 1-73.

3) E. Clarkson, K. Desai, and J. Foley, Resultmaps: Visualization for search interfaces, IEEE Transactions on Visualization and Computer Graphics, 15(6) (2009) 1057-1064.

4) M. Bruls, K. Huizing and J. van Wijk, Squarified treemaps, Data Visualization 2000 (2000) 33-42.

5) W. Chen, H. Wynne and L. L. Mong, Tagcloud-based explanation with feedback for recommender systems, Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (2013) 945-948.

6) J. Steele, L Noah, Beautiful visualization (O'Reilly Media Inc, 2010) 40-60.

7) S. Molnar, M. Robert and B. Maria, Trending Words in Digital Library for Term Cloud-based Navigation, In SMAP '13: Proc. of the 8th Int. Workshop on Semantic and Social Media Adaptation and Personalization (2013) 1-4.

8) B. Tessem, J. Bjarte and V. Csaba, Mobile Location-Driven Associative Search in DBpedia with Tag Clouds, I-SEMANTICS 2013 Posters and Demos (2013) 6-10.

9) J. Wang, J. Zhao, S. Guo, C. North, Clustered Layout Word Cloud for User Generated Review (2012) 1-10.

10) S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: A Nucleus for a Web of Open Data, The Semantic Web (2007) 722-735.

11) HMS. Wen, GH. Eshley and F. Bond, Using WordNet to predict numeral classifiers in Chinese and Japanese, GWC 2012 6th International Global Wordnet Conference (2012) 211.

12) 「ユーザビリティハンドブック編集委員会」, ユーザビリティハンドブック (共立出版, 2007).

13) 内閣官房ＩＴ担当室, 電子政府ユーザビリティガイドライン (2009).

14) H. B. Mann and D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, The annals of mathematical statistics, 18 (1947) 50-60.

15) G. Salton and M. J. McGill, Introduction to modern information retrieval (1986).

16) H. Rosling, Gapminder: http://www.gapminder.org (2009).

付 録A　付録A 基礎技術 (WORDNET)

ここでは入力キーワードとタグの意味的類似度を測定する際に用いる概念辞書 WordNet について述べる． Wordnet は単語の上位 / 下位関係, 部分 / 全体関係, 同義関係, 類義関係などによって単語を分類し体系づけた, 英語の概念辞書である [11)]. WordNet では英単語が synset と呼ばれる同義語のグループに分類され, 簡単な定義や, 他の同義語のグループとの関係が記述されている. WordNet の目的は直感的に使うことのできる辞書とシソーラスが組み合わされた成果物を作ること, および自動的文書解析や人工知能のアプリケーションの実現を支援することにある. この WordNet から着想を得て開発されたもので日本語 WordNet がある. 独立行政法人情報通信研究機構（NICT）が開発したもので, 大規模かつ誰でも利用できる日本語の意味辞書である. 日本語 WordNet は, 語を synset でグループ化している点に特徴があり, 一つの synset が一つの概念に対応している. また, 各 synset は上位下位関係などの多様な関係で結ばれている. 日本語ワードネットに収録された synset 数や単語数, 語義数は以下のとおりである.

57,238 概念 (synset 数),93,834 語,158058 語義 (synset と単語のペア),135,692 定義文,48,276 例文

日本語辞書データベースの中で, 日本語 WordNet は無料で使用でき, 誰でも開発可能であるということから提案システムでは WordNet を採用した.

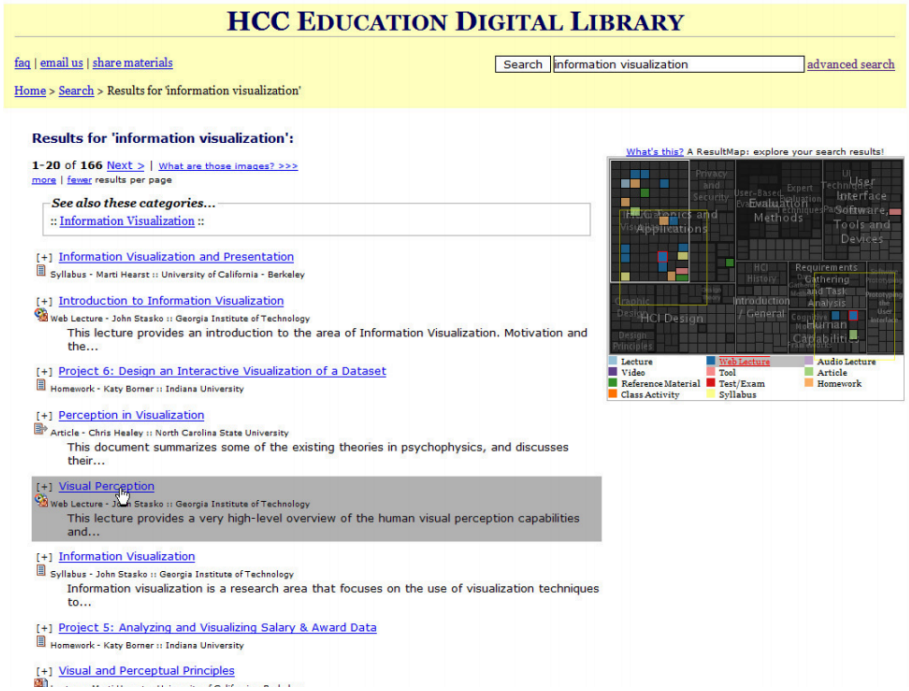


図 2.1: ResultMaps

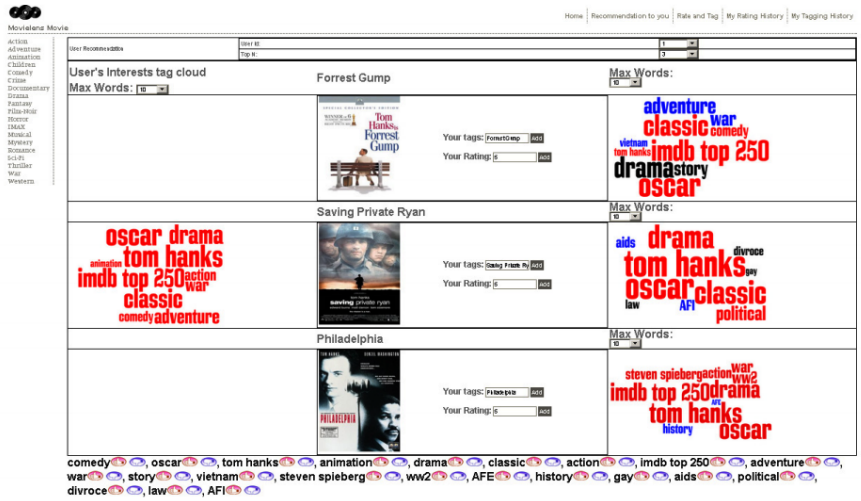


図 2.2: W. Chen の提案システム

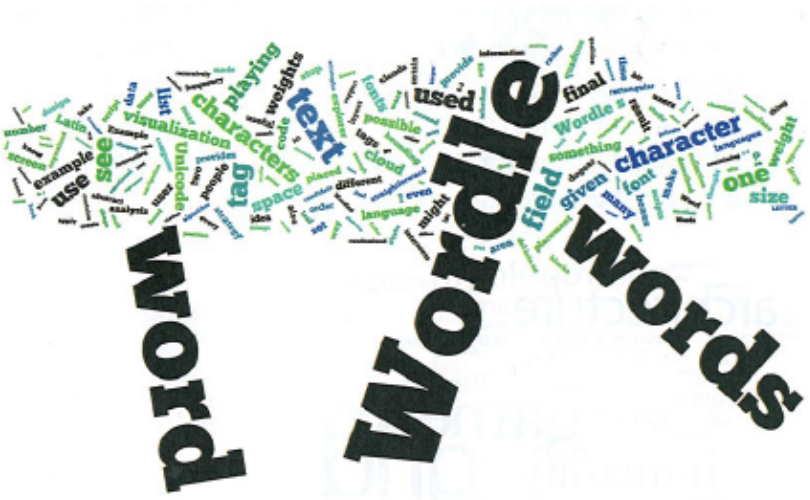


図 2.3: Wordle

図 2.4: B. Tessem の提案システム

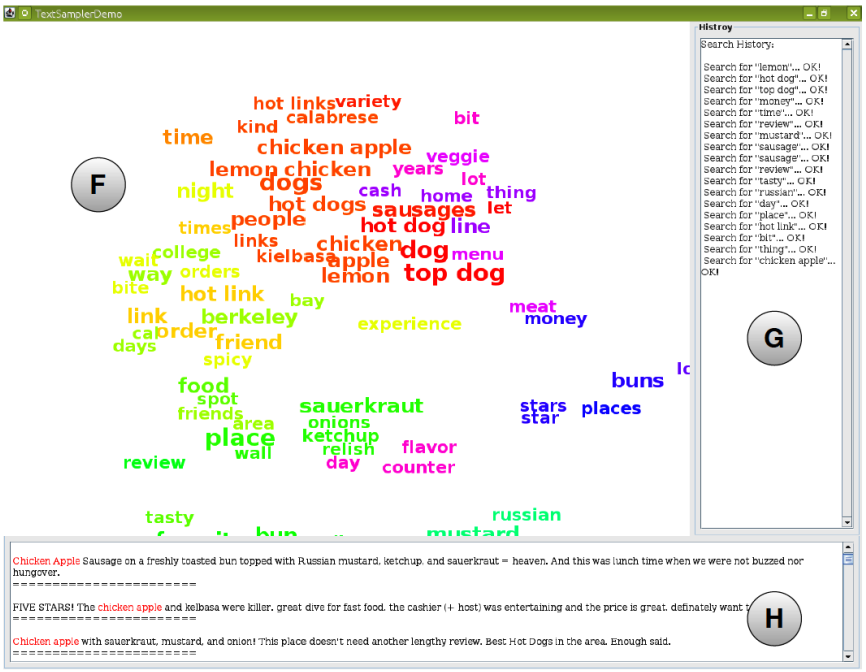


図 2.5: Ji. Wang の提案システム

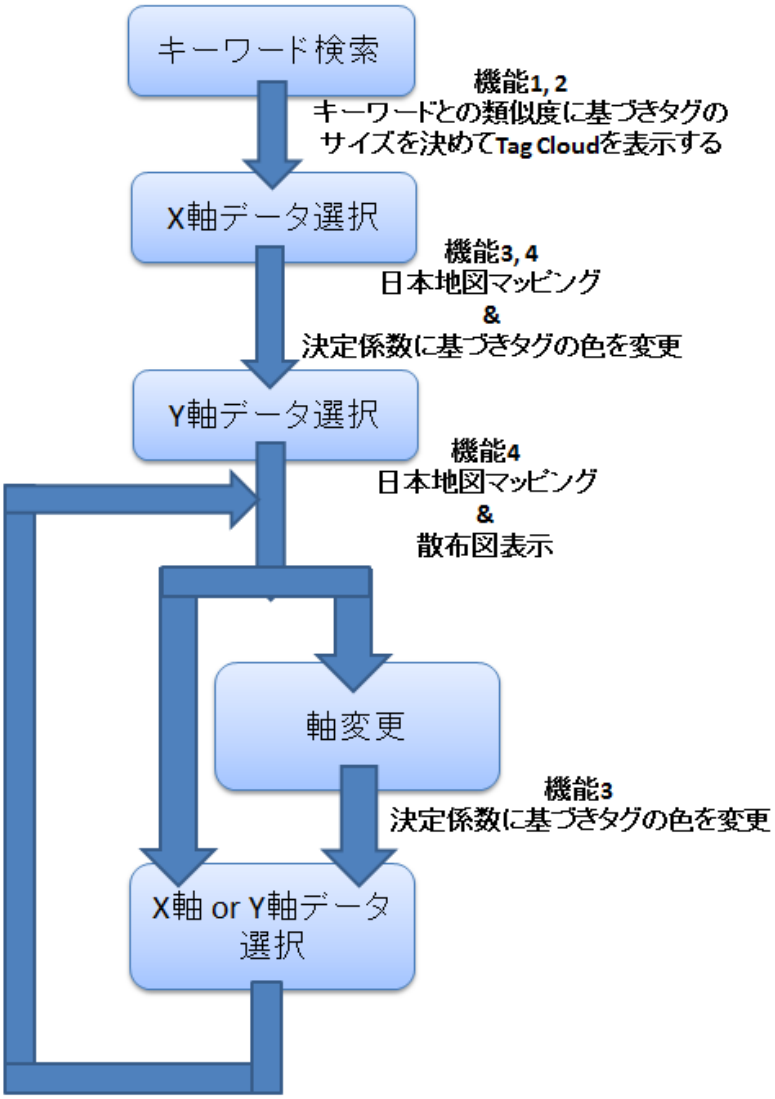


図 3.1: システム利用の流れ

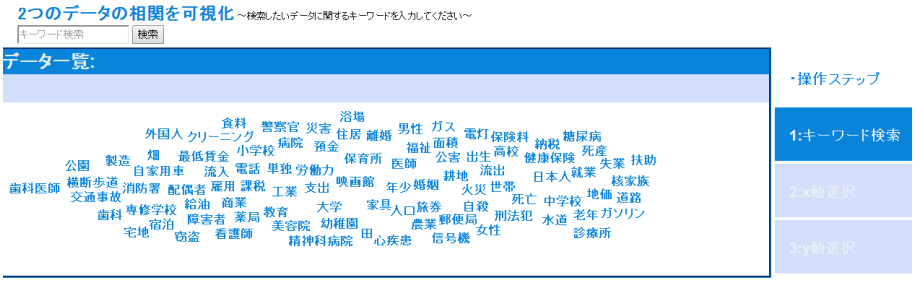


図 3.2: 提案システムの実行例 1

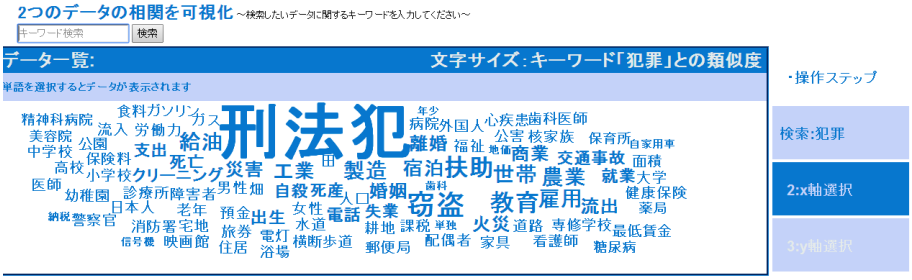


図 3.3: 提案システムの実行例 2

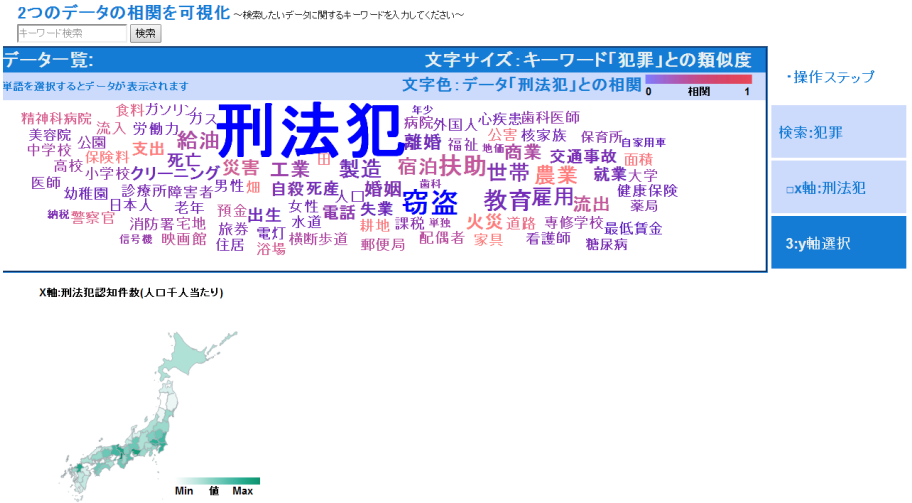


図 3.4: 提案システムの実行例 3

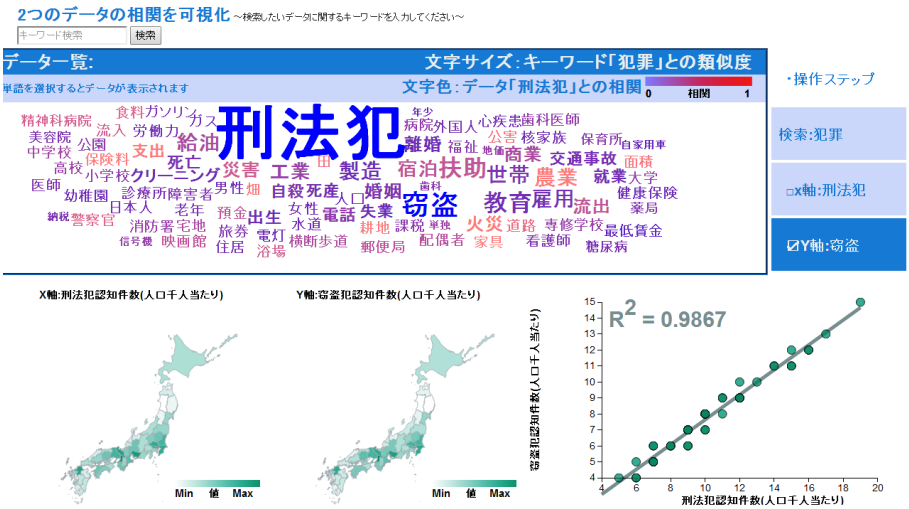


図 3.5: 提案システムの実行例 4

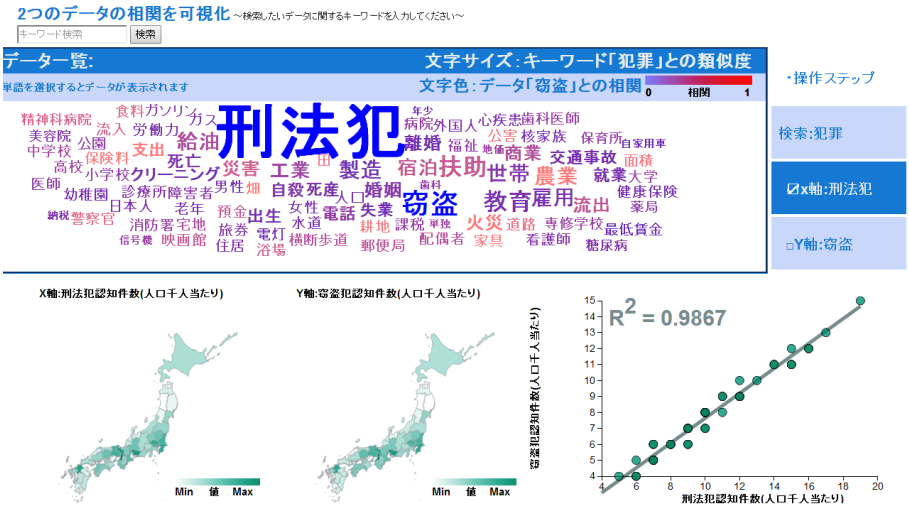
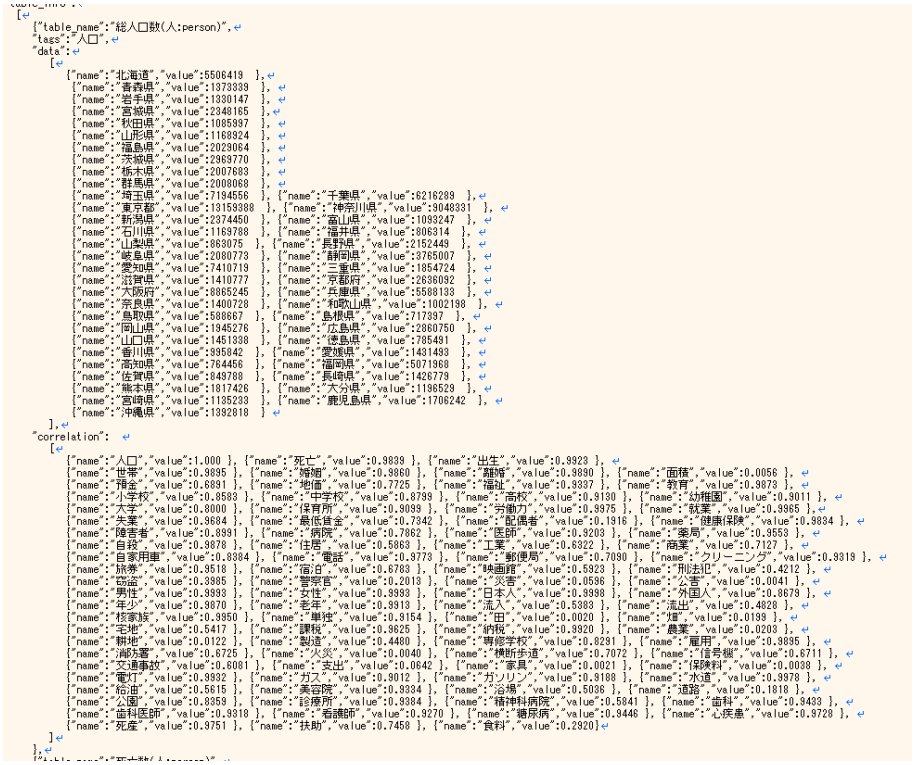


図 3.6: 提案システムの実行例 5



```
[
  {"table_name":"総人口数(人:person)",
   "tags":"人口",
   "data":
     [
       {"name":"北海道","value":5506419 },
       {"name":"青森県","value":1373339 },
       {"name":"岩手県","value":1330147 },
       {"name":"宮城県","value":2348165 },
       {"name":"秋田県","value":1085997 },
       {"name":"山形県","value":1168924 },
       {"name":"福島県","value":2029064 },
       {"name":"茨城県","value":2969770 },
       {"name":"栃木県","value":2007683 },
       {"name":"群馬県","value":2008068 },
       {"name":"埼玉県","value":7194556 }, {"name":"千葉県","value":6216289 },
       {"name":"東京都","value":13159388 }, {"name":"神奈川県","value":9048331 },
       {"name":"新潟県","value":2374450 }, {"name":"富山県","value":1093247 },
       {"name":"石川県","value":1169788 }, {"name":"福井県","value":806314 },
       {"name":"山梨県","value":863075 }, {"name":"長野県","value":2152449 },
       {"name":"岐阜県","value":2080773 }, {"name":"静岡県","value":3765007 },
       {"name":"愛知県","value":7410719 }, {"name":"三重県","value":1854724 },
       {"name":"滋賀県","value":1410777 }, {"name":"京都府","value":2636092 },
       {"name":"大阪府","value":8865245 }, {"name":"兵庫県","value":5588133 },
       {"name":"奈良県","value":1400728 }, {"name":"和歌山県","value":1002198 },
       {"name":"鳥取県","value":588667 }, {"name":"島根県","value":717397 },
       {"name":"岡山県","value":1945276 }, {"name":"広島県","value":2860750 },
       {"name":"山口県","value":1451338 }, {"name":"徳島県","value":785491 },
       {"name":"香川県","value":985842 }, {"name":"愛媛県","value":1431493 },
       {"name":"高知県","value":764456 }, {"name":"福岡県","value":5071968 },
       {"name":"佐賀県","value":849788 }, {"name":"長崎県","value":1426779 },
       {"name":"熊本県","value":1817426 }, {"name":"大分県","value":1196529 },
       {"name":"宮崎県","value":1135233 }, {"name":"鹿児島県","value":1706242 },
       {"name":"沖縄県","value":1392818 }
     ],
   "correlation":
     [
       {"name":"人口","value":1.000 }, {"name":"死亡","value":0.9839 }, {"name":"出生","value":0.8923 },
       {"name":"世帯","value":0.9895 }, {"name":"婚姻","value":0.9860 }, {"name":"離婚","value":0.9890 }, {"name":"面積","value":0.0056 },
       {"name":"預金","value":0.6891 }, {"name":"地価","value":0.7725 }, {"name":"福祉","value":0.9337 }, {"name":"教育","value":0.9873 },
       {"name":"小学校","value":0.8583 }, {"name":"中学校","value":0.8799 }, {"name":"高校","value":0.9130 }, {"name":"幼稚園","value":0.9011 },
       {"name":"大学","value":0.8000 }, {"name":"保育所","value":0.9099 }, {"name":"労働力","value":0.9975 }, {"name":"就業","value":0.9965 },
       {"name":"失業","value":0.9604 }, {"name":"最低賃金","value":0.7342 }, {"name":"配偶者","value":0.1316 }, {"name":"健康保険","value":0.9834 },
       {"name":"障害者","value":0.8991 }, {"name":"病院","value":0.7862 }, {"name":"医師","value":0.9203 }, {"name":"薬局","value":0.9553 },
       {"name":"自殺","value":0.9878 }, {"name":"住居","value":0.5863 }, {"name":"工業","value":0.6322 }, {"name":"商業","value":0.7127 },
       {"name":"自家用車","value":0.8384 }, {"name":"電話","value":0.9773 }, {"name":"郵便局","value":0.7090 }, {"name":"クリーニング","value":0.9319 },
       {"name":"旅券","value":0.9518 }, {"name":"宿泊","value":0.6783 }, {"name":"映画館","value":0.5923 }, {"name":"刑法犯","value":0.4212 },
       {"name":"窃盗","value":0.3985 }, {"name":"警察官","value":0.2013 }, {"name":"災害","value":0.0596 }, {"name":"公害","value":0.0041 },
       {"name":"男性","value":0.9893 }, {"name":"女性","value":0.9893 }, {"name":"日本人","value":0.9898 }, {"name":"外国人","value":0.8679 },
       {"name":"年少","value":0.9870 }, {"name":"老年","value":0.9918 }, {"name":"流入","value":0.5383 }, {"name":"流出","value":0.4028 },
       {"name":"核家族","value":0.9950 }, {"name":"単独","value":0.9154 }, {"name":"田","value":0.0020 }, {"name":"畑","value":0.0199 },
       {"name":"宅地","value":0.5417 }, {"name":"課税","value":0.9825 }, {"name":"納税","value":0.9920 }, {"name":"農業","value":0.0203 },
       {"name":"耕地","value":0.0122 }, {"name":"製造","value":0.4480 }, {"name":"専修学校","value":0.8291 }, {"name":"雇用","value":0.9895 },
       {"name":"消防署","value":0.6725 }, {"name":"火災","value":0.0040 }, {"name":"横断歩道","value":0.7072 }, {"name":"信号機","value":0.6711 },
       {"name":"交通事故","value":0.6081 }, {"name":"支出","value":0.0642 }, {"name":"家具","value":0.0021 }, {"name":"保険料","value":0.0038 },
       {"name":"電灯","value":0.9832 }, {"name":"ガス","value":0.9012 }, {"name":"ガソリン","value":0.9188 }, {"name":"水道","value":0.9878 },
       {"name":"給油","value":0.5615 }, {"name":"美容院","value":0.9334 }, {"name":"浴場","value":0.5036 }, {"name":"道路","value":0.1818 },
       {"name":"公園","value":0.8959 }, {"name":"診療所","value":0.9384 }, {"name":"精神科病院","value":0.5841 }, {"name":"歯科","value":0.9433 },
       {"name":"歯科医師","value":0.9318 }, {"name":"看護師","value":0.9270 }, {"name":"糖尿病","value":0.9446 }, {"name":"心疾患","value":0.9728 },
       {"name":"死産","value":0.9751 }, {"name":"扶助","value":0.7458 }, {"name":"食料","value":0.2920}
     ]
  },
```


図 3.7: 使用データ (json)

データリスト					
	(データ)			(データ)	
table_name(データ名)	総人口数(人:person)			食料費[二人以上の世帯](円:yen)	
tags(タグ)	人口			食糧	
	name(データ値名)	北海道		name(データ値名)	北海道
	value(データ値)	5506419		value(データ値)	64746
	name(データ値名)	青森県		name(データ値名)	青森県
	value(データ値)	1373339		value(データ値)	64016
data(データ値集合)	⋮	⋮		⋮	⋮
	name(データ値名)	鹿児島県		name(データ値名)	鹿児島県
	value(データ値)	1706242		value(データ値)	60967
	name(データ値名)	沖縄県		name(データ値名)	沖縄県
	value(データ値)	1392818		value(データ値)	54297
	name(データ値名)	人口		name(データ値名)	人口
	value(データ値)	1.0000		value(データ値)	0.2920
	name(データ値名)	死亡		name(データ値名)	死亡
	value(データ値)	0.9839		value(データ値)	0.2506
correlation(決定係数)	⋮	⋮		⋮	⋮
	name(データ値名)	扶助		name(データ値名)	扶助
	value(データ値)	0.7458		value(データ値)	0.0832
	name(データ値名)	食糧		name(データ値名)	食糧
	value(データ値)	0.2920		value(データ値)	1.0000

図 3.8: データ構造説明

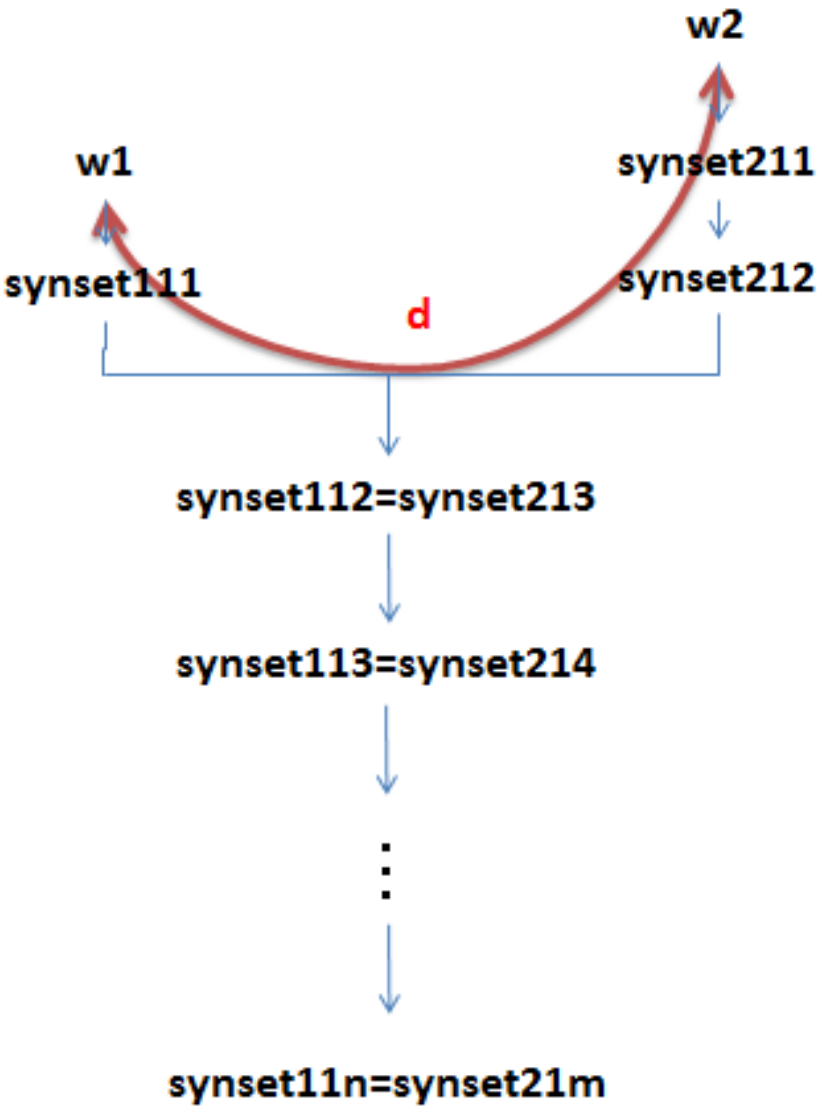


図 3.9: WORDNET のツリー構造

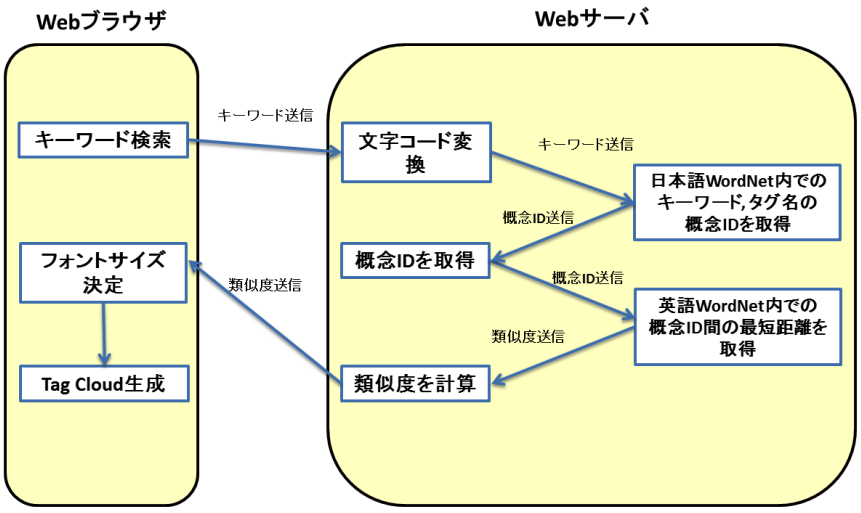


図 3.10: キーワード検索を行ってから Tag Cloud を生成するまでの流れ

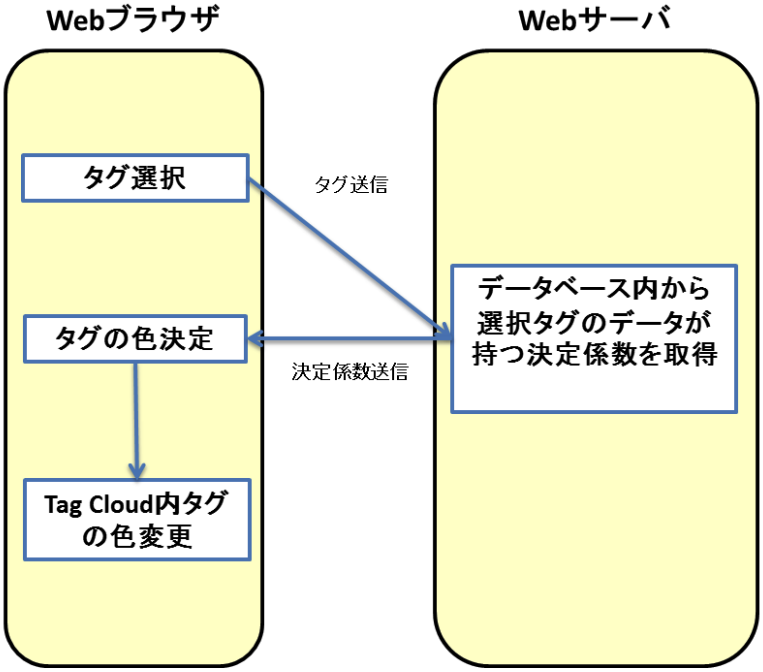


図 3.11: タグ選択からタグの色が変更するまでの流れ

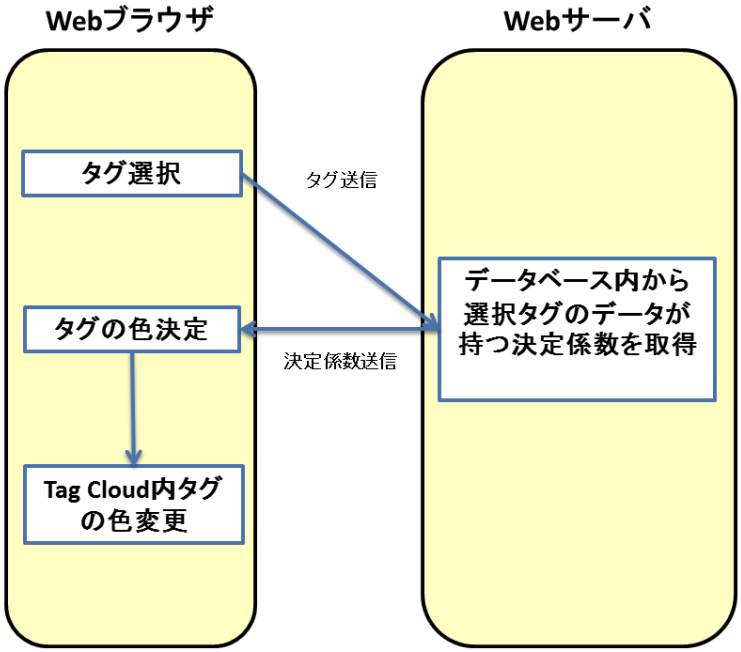


図 3.12: タグ選択から日本地図, 散布図, 回帰直線, 決定係数表示までの流れ

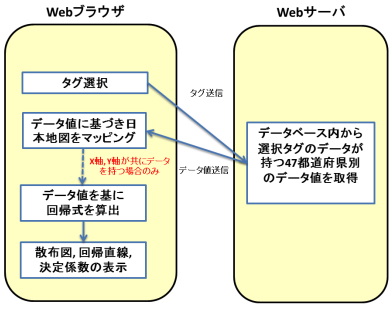


図 3.13: 軸変更からタグの色が変更するまでの流れ

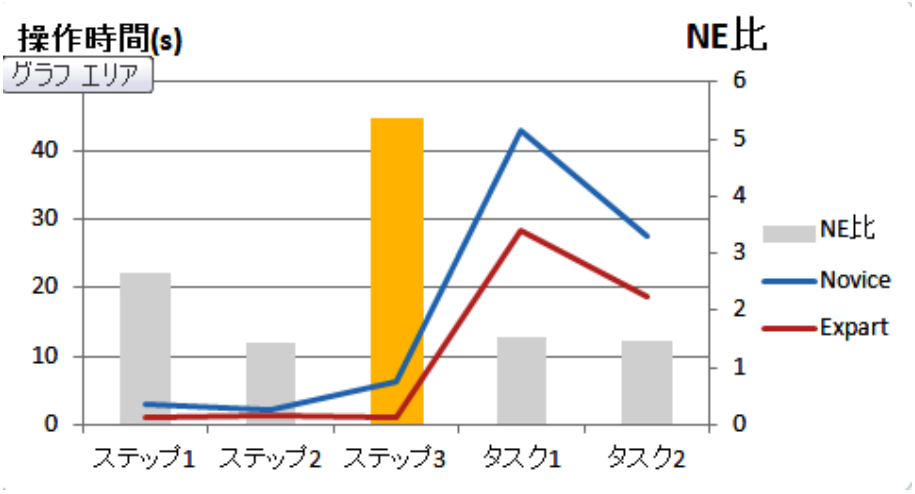


図 4.1: NEM(操作性能測定) 実験結果

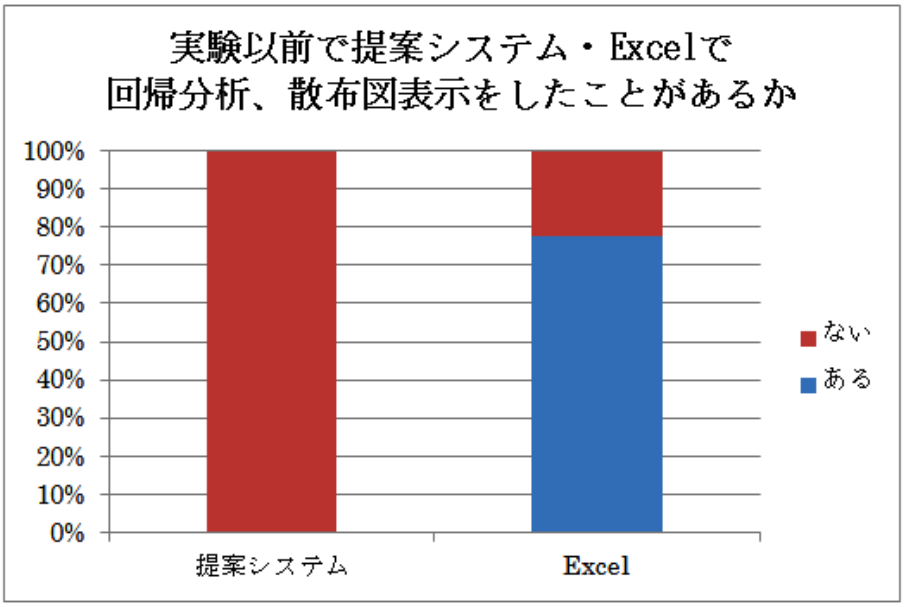


図 4.2: 対比実験実験結果

表 1: 対比実験結果

	提案システムの平均	Excel の平均	P 値
システムの使いやすさ	4.27	3.82	0.013
データの発見しやすさ	3.67	2.93	0.012
実験に要した時間 (min)	16.95	19.91	0.050
散布図生成/時間 (個/min)	0.42	0.29	0.022

表 2: ユーザーアンケート結果

類似度検索は仮説検証に役だったか	3.890
類似度をフォントサイズで表示したのは分かりやすいか	4.103
類似度計算にかかる時間	3.363
類似度計算ができない単語はどれくらいでましたか	3.246
相関を文字の色で表示したのはわかりやすかったか	4.129
データタグ (文字) の配色は適切でしたか	3.623
回帰直線は仮説検証に役だったか	4.342
決定係数は仮説検証に役だったか	4.150
散布図は仮説検証に役だったか	4.136
日本地図は仮説検証に役だったか	3.520
タグクリックでの軸変更はすぐわかりましたか	3.082
軸変更ボタンの配置は適切だと思いますか	3.714