

## 時間情報を持つテキストの 年表化、インタラクティブ 年表と年表マークアップ

松原 伸人

matubara@sra.co.jp

### ◆はじめに

時間情報を持つ大量のテキストの特徴を、インタラクティブに見ていくシステムを研究開発しています。

年表は出来事を分類して時間順に並べたタイムラインを並べて提示する表示方法です。データの時間的な変化を見ていて、一方のデータの変化を見つけたときに、ほかのデータがどうなっているかを見て、データがあるから関係がありそうだとか、歴史年表を見て、この時の事件はこういう社会状況がきっかけで起こったのかと想像したり、推測を促したり、想像を働かせたりすることができる道具と考えることができます。

人工知能、テキストマイニングやデータクラスタリングを用いて、データの特徴抽出、分類が機械的に高精度で高速に行えるようになってきています。

人が分類する場合と機械的な分類の違いとして、分類の意味が人によって決まっているかどうかがあります。

人が分類する場合、いくつに分類するか、どういう意味で分けるかということを決めてから分類したり、分類しながら意味が決まって行ったりするので、最終的

に出来上がる分類結果の意味を理解している状態になっています。

機械分類の場合、コンピュータが提示する分類結果を見て、各分類がどういう意味なのかを読み解かなければなりません。分類の要約や、ラベリングなども行われていますが、最終的には人間が結果を見て判断する必要があります。

分類結果を見て行く時に、時間順に読み進めることで理解しやすくなる、他の分類結果や付随する他のデータと見比べながら読むことで意味を拾い出しやすくなると思っています。

### ◆年表を作る

図1は2016年9月7日から3日間、“京都大学サマーデザインスクール” [1] の中で行ったワークショップ“リアルタイムな観察記録表現伝達のクロニクル” [2] で作成したワークショップ年表です。

ワークショップ参加者3名が観察者となり、他の1つワークショップ参加者の行動をFastEver2で記録、Excelで編纂して年表を作成しました。

観察者は各自の視点で見た事実をテキストで入力してFastEver2で投稿していきました。

二日目までは観察記録を中心に進め、三日目の午前中まで編纂と発表スライドの作成を行いました。

3名の観察記録は、記録時間順に1つのCSV形式で記録されていきました。

編纂によって、文言を統一、適切な語や言い回しに編集、出来事を3つの視点から分類しました。

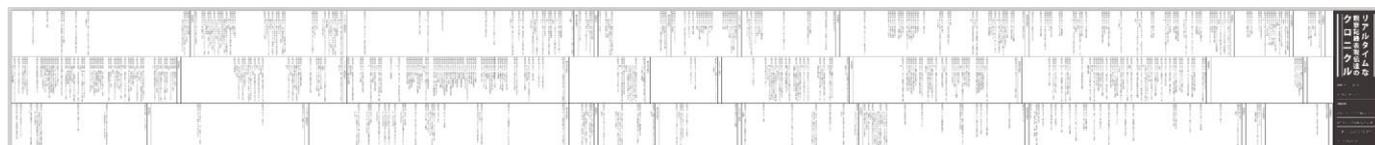


図1 ワークショップで作成した年表

[1] 京都大学サマーデザインスクール 2016 <http://www.design.kyoto-u.ac.jp/sds2016/>

[2] リアルタイムな観察記録表現伝達のクロニクル <http://www.design.kyoto-u.ac.jp/sds2016/theme31.html>

私は実施者としてワークショップに参加し、観察者が作成する年表データを見て、年表作成プログラムのデータフォーマットを作成したり、観察者と一緒に記録の時間的な分類をつける作業を行ったり、他のワークショップ記録を表示して欲しいといったリクエストに応じるために、ひたすらプログラムを作成していました。

## ◆ GSLetterNeo 年表

GSLetterNeo の年表を作成し、付録としました。

GSLetterNeo は毎月 SRA で発行している技術紹介記事です。2008 年から毎月発行され今回でちょうど 100 号目を迎えました。ひと月に 1 号と数えているので例外的にひと月に 2 号発行されている時は Vol.79+ などになっていて、現在合計 104 号発行されています。

GSLetterNeo の発行日の各号には抽出した特徴語を 10 語ずつ記載しています。特徴語は、**Vol.97 “テキストデータ群の重ね合わせによるヒストリと特徴のインタラクティブ表示(3)”** で紹介した TFIDF を用いて抽出した、Vol.1 から Vol.99 までの各ドキュメントにおける各単語の出現頻度の高い単語です。

特徴語を抽出する際に計算した各号の語彙数と、発行日ごとの語彙数の合計をグラフにしています。

一般的なグラフでは時間軸の時間の刻みが一定だったり、対数グラフのようにある関数に基づいて決まっていたりしますが、この年表のグラフは、“GSLetterNeo の歴史” タイムラインに現れる出来事の位置に、その時発行された号の語彙数や、その時点での語彙の合計数をプロットするような仕組みになっています。

## ◆ GSLetterNeoWeb 年表

色々な年表を見ていると、“年表” の時間軸の刻みは、掲載する出来事の量によって決まることが多いことがわかります。

開発中の年表化プログラムは、CSV のようなシンプルな時系列の出来事データ群を読み込んで、時間ごとに出来事を分類して表示します。

各号が発行された同じ頃に、インターネット上でどのような関連記事が出ているかを見られるように年表にしました。SRA 先端技術研究所 (KTL) の GSLetterNeo のページに掲載しています。

<http://www.sra.co.jp/ctl/gletterneo/index.html>

関連記事は 2016 年 11 月時点での GSLetterNeo の各号から抽出した特徴語 10 語を用いて Google 検索した結果 10 件を用い、その 10 件の中から SRA のページと更新日時が無いページを省いて、時間順に並べてあります。

GSLetterNeo のタイトルをクリックすると、別のウインドウに PDF を表示します。

各号の特徴語群をクリックすると、その特徴語群に関連する記事をハイライトします。関連する記事を見つけて読んだり、思いもよらない関係なさそうな分野の記事が関係していることを見つけたり、昔書いた記事に関係するトピックが最近増えていたりなどといったことがわかりおもしろいです。

特徴語は、新しい号が発行されるたびに語彙が増えるため変化します。

インターネット上の Web ページは日々新たに増えたり、更新されたり、なくなったりします。

特徴語の変化、インターネットの変化によって年表にでてくる関連記事も変化していきます。この年表は、作成するタイミングによって異なる内容が出来上がることになります。

**GSLetterNeo の歴史を通して、変化するインターネットの瞬間をとらえ、関連する記事を記録した年表、**ということですが。

夢を。

GSLetterNeo Vol. 100  
2016 年 11 月 20 日発行  
発行者 ● 株式会社 SRA 先端技術研究所  
編集者 ● 土屋正人

バックナンバーを公開しています ● <http://www.sra.co.jp/gletter>  
ご感想・お問い合わせはこちらへお願いします ● [gsneo@sra.co.jp](mailto:gsneo@sra.co.jp)



**株式会社SRA**

〒171-8513 東京都豊島区南池袋 2-32-8

夢を。Yawaraka Innovation  
やわらかいのべしょん

# 付録

# GSLetterNeo年表

GSLetterNeo年表は、100号を記念して2016年11月に作成しました。

GSLetterNeoには11分野があります。

時代は、その期間に執筆されたドキュメントの傾向を表しています。

分野ごとにタイムラインがあり、発行した年月とタイトルと著者が出てきます。

年数を表す場合は太字を用いています。

Volは号数を表しています。

各号はドキュメントのタイトルと著者をスラッシュで区切って表しています。

各号の下に並ぶ単語は、全号本文を対象にTFiDFを用いて抽出した類似度の高い10単語であり、その号の特徴を表しています。

各号のキーワードはゴシック体の太字で表記しています。

語彙の変化には各号の語彙数と、その時点での全号の合計語彙数を載せました。

青字は各号の語彙数を表します。

緑字はその時点でのGSLetterNeo全号の合計語彙数を表します。

