212-NFC0-1

Total Environment for Text Data Mining

砂山 渡
Wataru Sunayama

広島市立大学 大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City Univeristy

In this challenge, an environment that we can combine plural text-mining techniques flexibly is constructed and is distributed widely. Text mining techniques include many topics such as key sentence extraction, keyword extraction, topic extraction, text coherence evaluation, multi-text summarization, text clustering and so on. However, such tools constructed by individual researchers exist separately, so people who want to use various techniques cannot activate them instantly. This environment preserves high motivation to keep studying, and many tools will be utilized in the world. People who want to concentrate on their creative activities can prepare and use a customized environment that consists of selected modules depend on their needs.

1. はじめに

本チャレンジでは，複数のテキストマイニング技術を柔軟に組み合わせて使える統合環境を構築し，電子テキストを扱う多くのユーザの，創造的活動を支援するツールの提供を目指す．

テキストマイニングと呼ばれる研究には,「重要文抽出」「キーワード抽出」「トピック抽出」「テキストの一貫性評価」「複数文書要約」「テキストクラスタリング」などさまざまな課題があり，すでに多くの研究成果も世の中で発表されてきている．しかし，それぞれの技術を利用するためのシステムやツールは，各研究者が独自に構築することが多く，また論文用の試験的なシステムとなっていたりするため，実際に世の中で使われる技術はごく一部に限られてしまっている．

また，情報を多角的に分析したいユーザは，複数のテキストマイニング技術を用いたいと考える．各研究者が配布用のシステムを提供していた場合でも，複数の技術を併用するためには，それらのシステムを各方面から別個に入手した上で，システム間のデータの受け渡しや結果の比較のために，手作業でフォーマットを整えたり，新たなインタフェースを独力で構築する必要が生じる．これらのことは，単に手間がかかるというだけでなく，直感的に試行錯誤を繰り返しながら知見を得る創造活動の妨げになる．

そこで，既存また将来の研究成果によるテキストマイニング技術を，１つのシステム内のモジュールとして扱うことができ，ユーザの選択したすべてのモジュールを連動して動作させられる環境を構築し，それを無償ツールとして公開することを目指す．これにより，先の問題点を解決する以下の効果が見込まれる．

- 各研究者が研究成果を一つのモジュールとして配付することを意識できるため，研究の高いモチベーションの維持につながること，また多くの技術の実用化や再利用が見込まれる．
- 複数の技術を用いたいユーザの環境が整えられ，ニーズに応じたモジュールを選択した上で，集中的して作業を行うことができる．

連絡先: 砂山渡，広島市立大学大学院情報科学研究科，731-3194
広島市安佐南区大塚東 3-4-1，TEL082-830-1705

2. チャレンジの詳細

本チャレンジでは，複数のテキストマイニング技術を柔軟に組み合わせて使える統合環境を構築し，作成した環境，および環境内で選択的に使用できるモジュールをダウンロードできる Web サイト (図 1) を立ち上げることを目指す．

2.1 環境の構成要素

作成する環境は，例えば，二次元ディスプレイ上の横 640pixel，縦 900pixel の領域を一つのモジュールのための領域として，この領域を縦と横に任意の数だけ並べることができるウインドウを作成する．現在の標準的なディスプレイであれば，1280 × 960pixel の解像度を出すことができるため，１画面内に２つの領域を表示できる．また 30 インチディスプレイ（解像度 2560 × 1600pixel）であれば，横に４つの領域 (図 2) を表示することができる．そのため，環境の構成要素には以下が挙げられる．

- 統合環境全体のウインドウ
- 環境を構成するモジュール群
- 環境とモジュールをつなぐインタフェース
- モジュールとモジュールをつなぐインタフェース

2.2 本チャレンジの計画

本チャレンジの５年間の計画を以下に示す．

- １年目：チャレンジ協力者募集，統合環境の仕様の決定
- ２年目：環境とモジュール間のインタフェースの決定，統合環境の構築
- ３年目：ダウンロードサイトの立ち上げ，統合環境と試験的基本モジュールのアップロード
- ４年目：モジュール間連動インタフェースの仕様決定，試験的連動モジュールのアップロード
- ５年目:ダウンロードサイトの正式運用開始，各種モジュールのアップロード受付開始

以下で，各計画の詳細について述べる．

図 1: 統合環境とモジュールのダウンロードサイト（イメージ）

2.2.1 環境の仕様の決定と構築

実際に統合環境を用いるユーザを想定し，環境全体のウインドウ構成を検討する．まず，環境内で用いられるモジュールの対象研究範囲を特定し，想定される入出力情報を列挙する．その後，各モジュールが要する領域のサイズとその配置の組み合わせ方法についての仕様を検討した上で決定する．環境及びモジュールの仕様言語としては，普及率が高く汎用的なオブジェクト指向言語として Java 言語を想定している．

また環境の機能として，モジュールの選択と配置，モジュール全体の初期化と再計算処理，テキストデータの入出力処理などのメソッド（関数）を定義した上で実装する．

2.2.2 環境とモジュール間のインタフェースの決定

統合環境とモジュール間でデータの受け渡しを行うためのインタフェースを定義する．すなわち環境側では，入力されたテキストデータに関する情報を保持する変数を定義し，それらへのアクセス方法を定義する．モジュール側では，環境内のモジュールとして動作するために実装しておくことが必要となる，モジュールの配置や大きさに関する変数や，モジュールの初期化を行えるメソッドを定義する．

また，Windows, Mac, Linux といった OS の違いや，日本語の文字コードの違い，英語や数値データへの適用可能性，などを踏まえたプラットフォームを検討する．

2.2.3 モジュール間のインタフェースの決定

環境内で使用する各モジュールは，それぞれが独立に動作するだけでなく，あるモジュール内での操作が，他のモジュールにも反映される仕組みを導入する．そのため，各モジュールが他のモジュールによる操作を認める際に，データの授受を行う変数やメソッドに関する仕様を定める．必要に応じて，統合環境が仲介する形でのモジュール間の連動となる可能性もある．

2.2.4 ダウンロードサイトの立ち上げ

統合環境と各種モジュールをダウンロードできるサイトを立ち上げる．Web サーバに環境とモジュールのアプリケーションを置き，ダウンロード環境を構築する．また研究者が，各自が作成したモジュールをアップロードできる CGI を実装する．

アップロードのためのアカウントの作成と管理，アップロードされたファイルのウイルスチェック，サーバのセキュリティ，著作権や免責事項など，運用面での仕様を定める．

3. テキストデータ分析のための統合環境

本章では，テキストデータを分析するための統合環境について述べる．

3.1 統合環境の利用目的

テキストデータを分析する局面において，以下のような統合環境の利用目的が挙げられる．

1. テキストデータの内容を詳細に理解したい
2. 客観的にテキストデータの内容を把握したい
3. より多くの知見をデータから獲得したい
4. 創造的活動において新たな行動戦略を創り出したい

これらの目的は，テキストマイニングの各システムの目的としても，挙げられる機会が多い．しかし，単一のシステムで達成できることは限られており，これらの目的を，十分に納得がいくレベルで達成するためには，複数システムの出力を組み合わせていくことが肝要と考えられる．

3.2 統合環境の特徴

統合環境では，複数のシステムからの出力を連動させる環境により，ユーザの操作の手間を省き，データ認識の効率を高めることができる．特に，ユーザ独自のモジュールの組合せにより，さまざまな目的に柔軟に対応でき，単純かつ直感的に，データからの気づきと多角的な分析を促すことができると考えている．

すなわち従来の，多くの既存システムの中から，目的に応じたシステムを選択した上で，複数のシステム間でデータのやり取りを必要とする状況に比べると，以下を本環境の特徴として挙げることができる．

- ユーザは目的に応じて，容易に独自のモジュールの組合せを設定して，使用することができる
- システムの入出力操作の回数を減らすことができる
- 複数のシステムからの出力を対応づけて眺められる
- あるシステムの視覚化インタフェース上で，複数のシステムへの入力を，一度に直感的に与えられる

これら直感的で単純な操作と，データ認識の効率化により，テキストデータの集中的な分析と，データからのアイデア発生を促すことが期待できる．

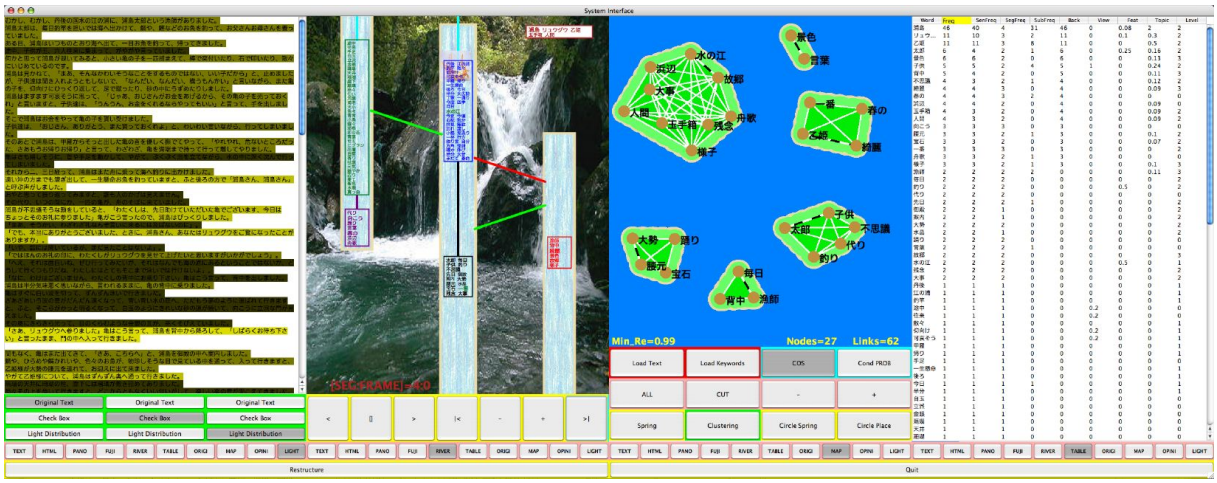

図 2: サンプル環境画面 (解像度 2560 × 1000)

3.3 サンプル環境

本節では，現在実装中のサンプル環境について述べる．モジュール群は主に，次の 2 つの目的によって大別される．

1. 特定のテキストについて，詳しい情報を知りたい
2. 複数のテキスト間の関係や情報を知りたい

入力となるテキストデータについて，1 つのテキストを入力する際には，その文の区切りを句点で，段落の区切りを特定のタグにより認識できる形式で入力する．複数のテキストを入力する際には，各テキスト内の文の区切りが句点で認識できる形式，また複数テキストを 1 つのテキストとして連結し，テキスト間の区切りが特定できるタグを挿入した上で入力する．

複数テキストを入力する際に 1 つのテキストとして連結するのは，ファイル入出力の回数を減らし実行時間の削減を図るとともに，テキスト間の区切りを，1 つのテキストを入力とした際の段落の区切りと同等に見なすことで，1 つのテキストを対象としたモジュールにも適用可能とするための措置となっている．

単一のテキストを入力として，そのテキストの情報を表示するサンプル環境上のモジュールを以下に示す．

1) テキスト表示（兼エディタ）
2) 単語の頻度情報の表示
3) キーワード表示
4) 要約表示 [相良 07]
5) テキストの一貫性表示 [砂山 08a]
6) 意見文表示 [川口 09]
7) 主題関連部分の表示 [西原 09]

複数のテキストを入力とし，そのテキスト間の関係を表示するサンプル環境上のモジュールを以下に示す．

8) 2 つのテキスト間の差分表示
9) クラスタリング結果表示 [Newman 04]
10) 独自性表示 [砂山 08b]
11) 具体抽象関係の表示 [砂山 09]（モジュールとして実装中）

図 2 に，パネル（各モジュールの表示領域）を 4 つ並べたサンプル環境の画面を示す（左から順にモジュール 7)，5)，9)，2)）．各モジュールは 640 × 900pixel の大きさのパネル上に表示することができ，横に並べるパネル数は，実行時引数により変えることができる．

ユーザは，使用したいモジュールを，環境下部のボタンを押すことで選択できる．現在，モジュール間の連動は一部でのみ実装されており，たとえば単一のテキストを入力した際に，モジュール 3) のキーワード表示モジュールにおいてキーワードを選択すると，モジュール 4) で選択されたキーワードを主題とした要約，モジュール 5) で選択されたキーワードを主題とした一貫性表示を行うことができる．

また，複数テキストとしてレポート集合を入力として与えたときに，モジュール 10) に加えて，3 つのテキスト表示パネルに，類似する 2 つのレポートをモジュール 1) として，その差分をモジュール 8) 上に連動させて表示することができる．代わりにモジュール 6) と連動させて，意見文が書かれているレポートとその独自性を比較することもできる．

4. チャレンジが 5 年以内に実現できる根拠

チャレンジを実現するための技術的な要素として，統合環境やモジュールを作成するためのプログラミング環境と，モジュールとなる既存技術の存在が挙げられる．

様々な要素技術を組み合わせて作成されるソフトウェアシステムでは，オブジェクト指向言語が一般に使われており，近年発達してきた Java 言語は，大学の授業にも取り入れられつつあるため，今後も広く普及して行くことが見込まれる．そこで，この Java 言語を利用することで，多くの研究者が統合環境の構築，またモジュールの構築に関わることができると考えられる．

また環境内で用いるための研究成果モジュールについて，テキストマイニングに関わる様々な技術が研究として発表されており，それら既存技術の中からいくつかの技術を初期モジュールとして作成することができると考えている．

本チャレンジの協力者が得られれば，2.2 で述べた計画に沿って進めて行くことにより，5 年以内の実現は可能と考えている．

5. 社会への貢献が期待できる根拠

現在の世の中は，多くの情報を獲得するとともに，それらをいかに分析して次の行動につなげていくかが問われている．その際に情報を多角的に分析できるツールは必須と考えられる．コンピュータを使っていて電子テキストを扱わない人はおらず，簡便で実用的な環境の上で，多角的にテキストを分析できるツールへのニーズと期待は高いと考えられる．

たとえば，新たに引き継いだ仕事の資料に目を通す際に，その量が膨大であった場合には，要点を抑えながら全体を把握したいと考える．そのためには，自動要約技術，キーワード抽出技術，トピック抽出技術など，複数の適用可能な技術が存在するが，あるユーザは，トピック抽出で得られたキーワードを主題とした，要約や関連キーワードを得たいと思うこともあり，各技術によるシステムを独立に動かしていただけでは，十分な理解につなげにくい場面も想定される．

自由記述アンケートの分析や，ブログ記事や掲示板からの流行の抽出など，これまでのテキストマイニングでも題材として扱われてきたような多くのニーズに対応して，キーワードを得るだけ，要約を得るだけ，トピックを抽出するだけ，ではなく，さまざまな情報を連動させて取り出して眺められる環境を本チャレンジで提供することによって，テキストデータの集中的な分析と総合的な判断を支援できると考えている．

また，既存のテキストマイニングのためのツールは，高価で販売されていたり，技術として確立されて時間が経過したものが含まれていることが多い．本チャレンジで構築する環境は無償で提供するため，個人が容易に入手可能なツールとなることや，最新の研究成果が，すぐに実世界で適用可能になることも期待できる．反面，モジュールによってはその信頼性に関する説明が十分でないものが現れる可能性もあるが，結果に対する最終的な判断を行うのはシステムではなく人間であり，リスクを認知して使い方を誤らなければ，最新の幅広い技術を用いた分析が可能になり，社会に大きく貢献できると考えている．

6. 人工知能への貢献が期待できる根拠

人工知能研究への貢献として，人工知能研究自体の促進，ならびに人工知能研究を世の中に広く認知してもらえることが期待できる．

前者に関しては，各研究者が研究成果を一つのモジュールとして配付することによる研究の実用化を意識できるため，研究の高いモチベーションの維持につながること，また他の研究との比較が容易になることが挙げられる．研究成果が論文として刊行されることは 1 つの研究のモチベーションとなり得るが，研究者としては，自分の作成したシステムが世の中で実際に使われて役に立つことを強く望んでいる．そのための環境が整うことは，研究を促進する上での大きな動機付けになると考えられる．

また，作成したシステムの評価を行う際には，関連システムとの比較が不可欠と考えられる．新しい技術が研究論文として発表されても，それが実際に活用されるためには，論文の著者からツールをもらうか，独力で実装する必要がある．1 つの共通の環境が存在して，そのモジュールとしてダウンロードが可能になれば，既存研究との比較も容易になると考えられる．

後者に関しては，実用的なツールが容易に入手可能な状況になれば，多くのユーザに使ってもらうことができ，世の中に広く認知されると考えられる．認知度が高まるにつれ，研究そのものへの関心も高まり，結果として共同研究や予算獲得などの面で研究環境が良くなることが期待できる．

7. チャレンジ協力者大募集

本チャレンジの実現のためには，環境の仕様の作成と実装，サイトの立ち上げなどに関連して，多くの協力者なくしてチャレンジは実現し得ないと考えている．提案者のプログラミングスキルやサイト運営のノウハウは不十分であるため，アドバイザ的なご協力はもちろんのこと，より積極的に本チャレンジを実現するためにご尽力頂けるコアメンバーも不可欠かつ募集中で，随時大歓迎となっている．

8. 結論

本チャレンジでは，複数のテキストマイニング技術を柔軟に組み合わせて使える環境を構築し，それらを広く提供することを目指している．本環境により，複数の技術を用いたいユーザの環境が整えられ，ニーズに応じたモジュールを選択した上で，集中的して作業を行うことができるようになると期待できる．

個別のテキストマイニング技術を開発する際に，他の技術との連携を意識して視野を広げつつ，本環境に統合することができれば，多くの研究が認知，実用化されるようになり，ユーザの情報の多角的な分析に基づく創造的開発，研究，経営戦略の立案などが支援されると期待できる．

提案環境の実現には，多くの方々のご賛同が必要と考えておりますため，様々な方面で，ご助言，ご助力賜れば幸いです．

参考文献

[西原 09] 西原陽子，佐藤圭太，砂山渡：光と影を用いたテキストのテーマ関連度の可視化, 人工知能学会論文誌，Vol.24, No.6, pp.480 – 488, (2009).

[相良 07] 相良直樹，砂山渡，谷内田正彦：サブトピックを考慮した重要文抽出による報知的要約生成, 電子情報通信学会論文誌, Vol.J90-D, No.2, pp.427 – 440, (2007).

[砂山 08a] 砂山渡：テキストの話の流れを視覚化するインタフェース, 第 22 回人工知能学会全国大会, 1B1-1, (2008).

[川口 09] 川口俊明，砂山渡：レポートのテーマ関連度と意見文抽出による情報量評価, 第 132 回情報処理学会ヒューマンコンピュータインタラクション研究会資料, pp.77 – 84, (2009).

[砂山 08b] 砂山渡，川口俊明：内容の独自性の視覚化によるレポートの独自性評価支援システム，人工知能学会論文誌，Vol.23, No.6, pp.392 – 401, (2008).

[砂山 09] 砂山渡，鮫島聡志，西原陽子：Web ページ間の相対的な具体抽象関係の視覚化による情報収集支援, 電子情報通信学会論文誌, Vol.J92-D, No.3, pp.271 – 280, (2009).

[Newman 04] M. E. J. Newman: Detecting community structure in networks, The European Physical Journal B, Vol. 38, No.2, pp. 321 – 330, (2004).