全体俯瞰分析を用いた着想支援とビッグデータ分析への応用

小野 謙二 [1),2)], 川鍋 友宏 [2)]

1) 国立大学法人九州大学 情報基盤研究開発センター
2) 国立研究開発法人理化学研究所 計算科学研究機構

keno@cc.kyushu-u.ac.jp

tkawanabe@riken.jp

An Idea Generation Support Using Bird's-Eye View Analysis and Its Application to Big Data Analysis

Kenji Ono [1),2)], Tomohiro Kawanabe [2)]

1) Research Institute for Information Technology(RIIT), Kyushu University.
2) Advanced Institute for Computational Science, RIKEN.

概要

ある１つの文書データ集合を対象に、切り口の異なる複数の分析結果を俯瞰的に表示し、そこから導かれる分析結果間の相関から、より深い関連性を把握することを意図し、全体俯瞰分析システムを構築した。システムは、データ収集部と分析部から構成され、ウェブ UI をもつ。検証のデータとして科学技術振興機構の運営する J-Global の論文・特許データベースを用い、ワードマップと論文・特許等の年別発行数を比べ合わせて分析することで、研究や商品開発のトレンドの推移と、そこに関わる研究者や関係機関の連携関係が可視化できることを確認した。

1 はじめに

計算機の性能向上とともに、その活用範囲は拡大し、スーパーコンピューターを用いた先端的なシミュレーションが生成するデータも大規模かつ多様なデータとなりつつある。 また、現実社会の課題解決のためには、多様なデータを総合的かつ俯瞰的に眺め、それらデータの背後にある意味を明らかにし、問題解決の糸口を探ることが期待されている。この場合、機械学習的なアプローチもあるが、人がもつ知識や洞察を利用することも劣らず有効であろう。

本論文では、検索対象とする文書空間におけるワードを、関連性を持たせて表示することにより、対象技術領域を俯瞰し、動向を把握することを狙いとしてシステム構築を行った。

2 全体俯瞰分析システム

2.1 システム概要

全体俯瞰分析システムとは、多次元視点分析による着想支援のシステムである。ある１つの文書データ集合を対象に、切り口の異なる複数の分析結果を俯瞰的に表示し、そこから導かれる分析結果間の相関から、より深い関連性を把握することを意図している。また、その結果から新たな情報検索条件の発掘や再検索により、セレンディピティ性の高い情報の発見も期待している。

図 1 は本システムの画面例である。ウェブブラウザ上で動作し、入力キーワードに基づいてデータベースを検索し、その結果から分析軸の違う各種グラフを表示し、それぞれ並べ替えや拡大・縮小が可能である。画面中段のツリー型グラフは「ワードマップ」と呼ばれ、本システムの中心的な分析グラフである。入力キーワードと相関の高い検索結果文書内の語彙について、その語彙間の相関係数に基づきツリー型グラフを構成している。ワードマップは表示面積の制約で同時表示ワード数を制限しているが、各ワードをクリックすることで、そのワードを中心とするワードマップを深掘りする再検索ができる。

2.2 セレンディピティ性

セレンディピティ (Serendipity) とは「*素敵な偶然に出会ったり、予想外のものを発見すること。また、何かを探しているときに、探しているものとは別の価値があるものを偶然見つけること。平たく言うと、ふとした偶然をきっかけに、幸運をつかみ取ること*」である[1]。本研究では「何かを探しているときに、探しているもの

とは別の価値があるものを偶然見つけること」をシステム的に支援し、それを研究やプロダクト開発の現場に応用することで、新奇性のあるアイデア創出のツールとなることを目的に本システムを開発した。

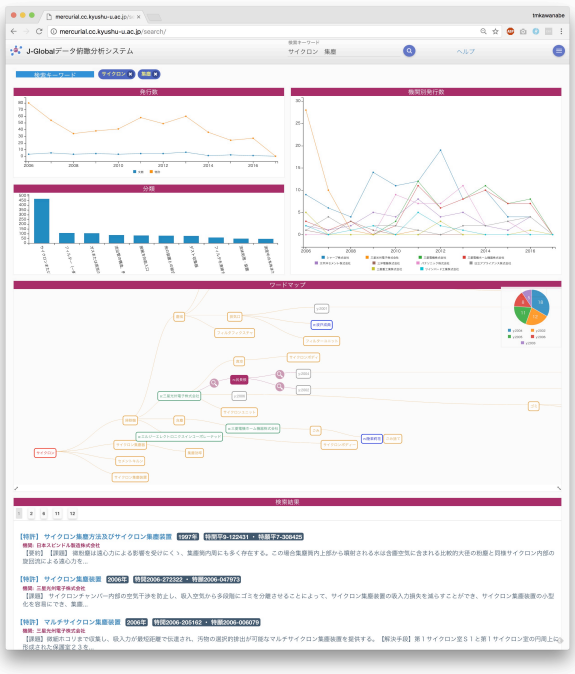

図 1. 全体俯瞰分析システム画面例

3 システム構成

本システムは、データ収集部と分析部から構成される。データ収集部は図 2 に示す機能群から構成されている。本稿では分析対象データソースとして、科学技術振興機構の運営する J-Global の論文・特許データベース[2]を用いる。

1. システム管理者は J-Global から収集したいトピックを収集ワードとして登録する。
2. クローラーは、J-Global WebAPI[3]を用いて検索結果を取得する。
3. 検索結果は主に文献（論文）、特許から構成されている。取得された個々の文書は、まずスクレイパーに渡され、JSON フォーマットで返されたデータをパースし、所定のデータ項目を抽出する(タイトル、概要、著者など)。
4. 形態素解析器(ChaSen[4])に入力し、文書を単語列に分解する。
5. 当該処理で得られた単語集合を文書単位で GETA に登録することで、単語の出現頻度を解析可能とする。

GETA[5] に登録されたデータはワードマップ向けのデータとして利用され、マイニングシステムを通して言葉の関係性を計算し、可視化結果を出力する。また、文献・特許の著者や発行年、カテゴリなどを リレーショナルデータベース へ登録することにより、発行件数の推移やカテゴリごとの集計分析が可能となる。検索対象データは HyperEstraie へ登録される。

図 3 は分析部の機能構成である。利用者が分析画面へアクセスし、検索語を入力することで分析処理が実行される。与えられた検索語は、インタラクティブ UI 部で関連語可視化処理、全文検索処理、サジェスト処理、集計分析処理等を実行する。それぞれの処理は、並列処理が可能となっており実際には、クライアントから個別のリクエストが非同期リクエストとし並行発行される。サーバ側では各機能に特化した view が個別の処理を返す。クライアントは応答が来た結果から順に可視化処理を実行し、分析結果がブラウザへ反映する。

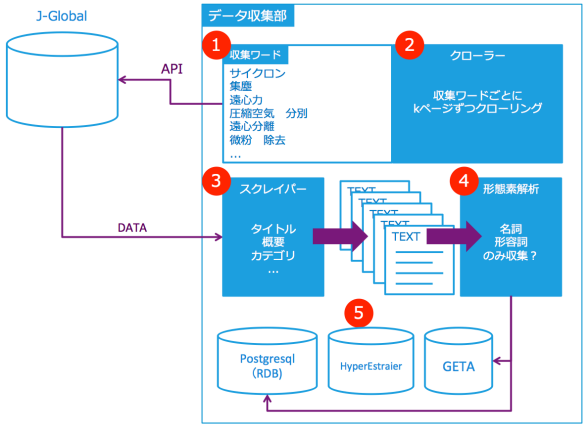

図 2. データ収集部機能の構成

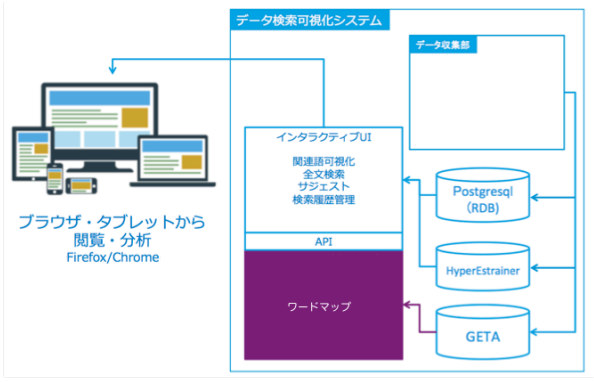

図 3. 分析部機能の構成

4 利用事例

本システムは内閣府戦略的イノベーション創造プログラム／革新的設計生産技術[6]における研究テーマ「全体俯瞰設計と製品設計着想支援」の一部として九州大学と理化学研究所が共同で開発し、その実証事例として J-Global から再生可能エネルギーに関するキーワード（風力、太陽光など）により抽出した約２０万件の文献および特許の文書情報を、本システムのデータソースとした分析処理を行った。ワードマップと論文・特許等の年別発行数を比べ合わせて分析することで、研究や商品開発のトレンドの推移と、そこに関わる研究者や関係機関の連携関係が可視化できることを確認した。

5 他のデータソースへの応用

前述のデータ収集部機能は、大別して、データの取得部分と、格納部分に分けられる。本稿で紹介した事例では、Web クローラが取得した文書データを、スクレイピング後に DB へ格納しているが、格納部分の処理のみを利用することで、Web を介さずに取得可能な静的な文書データを分析対象として取り扱うことも可能である。したがって、本システムを学内や企業内に蓄積された情報の知識ベースのプラットフォームとして応用可能な基本設計となっている。

また、近年利活用の発展が目覚ましい LOD(Linked Open Data)からセレンディピティ性のある知見を得るための分析手段として、本システムの応用が考えられる。本システムは構造化されたテキストデータを分析対象としており、その点でメタデータ構造が明確な LOD はデータソースとしては最適である。しかし、現在の本システムの実装は全ての文書情報を一箇所に集めた上でその相関を計算する仕組みであり、これを分散システムに対応するため実装の検討が今後の課題である。

6 まとめ

本稿では文書集合からセレンディピティ性のある発見を支援する「全体俯瞰分析システム」を紹介した。文書内に現れる特徴語の関連性を表したワードマップと呼ばれるグラフと、通常の統計量グラフを同時に俯瞰することで、探しているものとは別の価値があるものを偶然発見することを支援するシステムである。本システムは Web クロールでデータ収集するが、静的な文書集合にも対応可能であり、将来的には LOD のようなビッグデータ対応も検討している。

7 謝辞

本システムは内閣府戦略的イノベーション創造プログラム／革新的設計生産技術における研究テーマ「全体俯瞰設計と製品設計着想支援」の一部として開発され、分析対象データ事例として、科学技術振興機構の運営する J-Global を利用した。

また、本システムの中心技術であるワードマップの生成には、九州大学情報基盤研究開発センター廣川佐千男教授らによる「関連研究探索のための検索可視化システム」[7]、および情報処理振興事業協会（ＩＰＡ）が実施した「独創的情報技術育成事業」の研究成果である汎用連想検索エンジン GETA[5]の技術を応用している。本システムの研究開発にご協力頂いた関係各位に御礼申し上げます。

参考文献

[1] https://ja.wikipedia.org/wiki/%E3%82%BB%E3%83%AC%E3%83%B3%E3%83%87%E3%82%A3%E3%83%94%E3%83%86%E3%82%A3
[2] http://jglobal.jst.go.jp/
[3] http://jglobal.jst.go.jp/help/webapi/
[4] http://chasen-legacy.osdn.jp/
[5] http://geta.ex.nii.ac.jp/geta.html
[6] http://www8.cao.go.jp/cstp/gaiyo/sip/index.html
[7] http://doi.org/10.1241/johokanri.58.447
[8] http://geta.ex.nii.ac.jp/geta.html