

研究情報マイニング ～ 学術分野のビッグデータ活用可能性を探る ～

瀬川 修†

† 中部電力株式会社 エネルギー応用研究所

1 はじめに

我々は学術論文のアーカイブを「知識源」とみなし、特定分野のサーベイや発想支援のためのテキストマイニング技術について検討を行っている。本稿ではデジタル化された大量の学術論文データを入力とし、主題分析や著者・組織のネットワーク構造の分析、さらには「発想支援」に寄与する因果関係分析やキーワード連想など、ビッグデータの活用可能性について初期検討した結果を報告する。

2 文書構造解析

文書構造解析は、解析のレベルにより「表層的構造化」と「意味的構造化」などに分類される。今回は前者のテキストの表層的な手掛りから構成要素(タイトル、著者情報、概要、本文等)を同定する表層的構造化を行う。

扱うデータ型式であるが、近年論文のデジタルデータはPDFが主流となっていることからPDFを対象とする。ここでは、PDFと相性がよいデータ構造としてXMLの中間形式を介し、文書構造解析を行う方式を採用した。XML構造に変換することにより、文字列の位置情報やフォントサイズ等が取得できるため、ルールベースの表層的構造化の解析に適している。

3 主題分析

ここで述べる「主題分析」とは、論文テキストから技術トピックを表す特徴的なキーワードを抽出する手法のことである。具体的には、下記手順によりキーワード抽出を行う。

1. 技術トピックに対する言及に関連の深い「手掛り表現」を定義し、係り受け解析により手掛り表現に係る特定品詞の単語(または複合語)を抽出する。
2. 上記で抽出したキーワードを重要度(TF-IDF値)でスコアリングする。

今回は表1に示す手掛り表現を用いた。係り受け解析は独自実装によるルールベース方式の解析器を用いた。

Research Information Mining - Utilization Possibilities of Bigdata in Academic Fields -

† Osamu Segawa (Segawa.Osamu@chuden.co.jp)
Chubu Electric Power Co.,Inc. (†)

表 1: 主題分析の手掛り表現の例

研究、調査、検討、提案、適用、評価、構築 試作、実装、実験、方式、方法、手法、影響
--

4 著者・組織のネットワーク分析

論文の「共著関係」に基づき、特定分野の学術的つながりを分析する手法を検討した。共著関係は研究者相互の結び付きを最もよく表す人的・組織的なリンク構造と考えられる。提案方式では、前述の文書構造解析で抽出した著者および組織名称を用い、第一著者の重要度を勘案した共著関係の有向グラフを論文ごとに生成し、アーカイブ全体でグラフを合成する。

共著関係の有向グラフ生成の例を図1に示す。

著者の掲載順 {1st, 2nd, 3rd, 4th} の場合



[グラフ生成例]

論文1の著者の掲載順 {1, 2, 3}
論文2の著者の掲載順 {2, 1, 3}
論文3の著者の掲載順 {3, 2}
論文4の著者の掲載順 {4} (単著)
{ }内の数字は著者ID

合成した有向グラフ

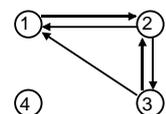


図 1: 共著関係の有向グラフの例

5 発想支援

ここでは研究の「発想支援」のアプローチとして、キーワードベースの2つの支援機能を提案する。

5.1 因果関係分析

因果関係分析は様々な手法 [1] が提案されているが、ここでは表層の「手掛り表現」に基づく簡易な手法を用いた。具体的には、「AによるB」、「Aが原因でB」など、あらかじめ定義した手掛り表現の左右に出現する特定品詞の単語(または複合語)を抽出することにより、因果関係(要因と結果)に係わるキーワードペアを同定する。さらに、論文ごとに上記手法で抽出した2項関係をアーカイブ全体で合成することにより、特定分野のキーワードの多段の因果関係を見出すことができると考えられる。

5.2 キーワード連想

ここでは文中の単語共起の指標として並立助詞に着目し、「名詞(または複合語)A」+「並立助詞」+「名詞(または複合語)B」というパターンに合致するキーワードペア A、B を連想関係にあると見なす。さらに、論文ごとに上記手法で抽出した 2 項関係をアーカイブ全体で合成することにより、特定分野のキーワードの多段の連想関係を見出すことができると考えられる。

6 分析例と考察

我々が開発を進めているテキストマイニングシステムをベースに前述の分析機能の実装を行い、電気学会全国大会の論文集(発表件数: 約 1500 件/年)を用いた分析を試行した。

6.1 主題分析

分野 3(情報処理、エレクトロニクス)の主題分析の例を表 2 に示す。トピック変遷を見ると年度ごとに特徴的なキーワードが出現していることがわかる。

表 2: 主題分析の例(電気学会全国大会)

(2012 全国大会 - 分野 3 上位 18 語) システム、センサ、ロボット、制御、評価、特性検討、利用、モデル、通信、認識、進化プラットフォーム、匂い、水素、IT、解析、開発
(2013 全国大会 - 分野 3 上位 18 語) システム、制御、センサ、エネルギー、デバイス温度、利用、効果、活性、運転、MEMS、開発計測、評価、最適、素子、形状、依存
(2014 全国大会 - 分野 3 上位 18 語) センサ、表面、手法、システム、モデル、作製加速度、アレイ、測定、提案、評価、技術、応答電極、体、データ、デバイス、管理

6.2 著者・組織のネットワーク分析

著者・組織のネットワーク分析の例を図 2 に示す。「出リンク」の多い著者がキーパーソン、「入リンク」の多い著者ほどアクティビティが高いと推定される。

6.3 因果関係分析

因果関係分析の例を図 3 に示す。アーカイブ全体での合成により多段の因果関係が得られている。

6.4 キーワード連想

キーワード連想の例を図 4 に示す。アーカイブ全体での合成により多段の連想関係が得られている。

7 関連研究

学术论文のテキストマイニングについては、様々な検討が行われている [2]。学会主導のプロジェクトとしては電子情報通信学会の論文データベースを用いた「I-Scover プロジェクト」[3] などがある。

8 まとめ

本稿では、学术论文のアーカイブを「知識源」として利用するためのテキストマイニング技術について検討し、ビッグデータ活用の見通しを得た。

参考文献

- [1] 坂地, 増山: “新聞記事からの因果関係を含む文の抽出手法”, 信学論 D, Vol.J94-D, No.8, pp.1496-1506, 2011.
- [2] 那須川, 西山, 吉田: “学术论文のテキストマイニング”, 言語処理学会 第 20 回年次大会, pp.800-803, 2014.
- [3] I-Scover URL: <http://iscover-p.ieice.org>

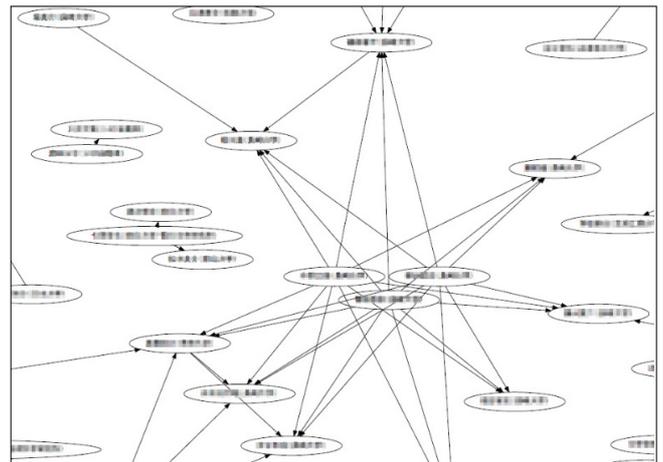


図 2: 著者・組織のネットワーク分析の例(電気学会 2014 全国大会論文集より生成したグラフ構造の一部)

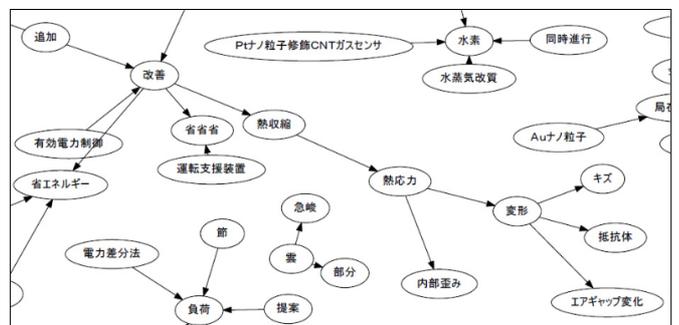


図 3: 因果関係分析の例(電気学会 2014 全国大会論文集より生成したグラフ構造の一部)

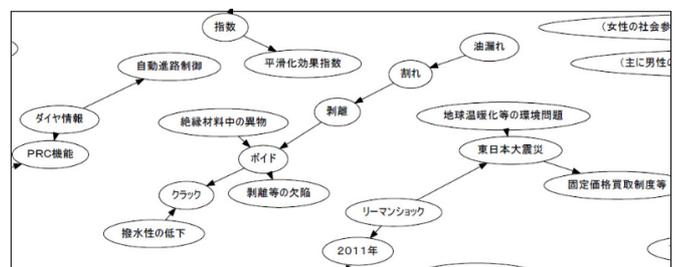


図 4: キーワード連想の例(電気学会 2014 全国大会論文集より生成したグラフ構造の一部)