

Title	テキストマイニングとネットワーク分析を用いた大規模文書群内の効果的知識抽出：知識構造マップによる知識構造の視覚化
Author(s)	松井, 亮佑; 妹尾, 大; 川島, 啓
Citation	年次学術大会講演要旨集, 24: 176-179
Issue Date	2009-10-24
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/8605
Rights	本著作物は研究・技術計画学会の許可のもとに掲載するものです。This material is posted here with permission of the Japan Society for Science Policy and Research Management.
Description	一般講演要旨

1 E O 2

テキストマイニングとネットワーク分析を用いた大規模文書群内の 効果的知識抽出—知識構造マップによる知識構造の視覚化—

○松井亮佑（株式会社ベイカレント・コンサルティング），
妹尾大（東京工業大学大学院），川島啓（財団法人未来工学研究所）

[要旨]

21 世紀を迎えて知識社会への移行が謳われる現在、政策や研究、ビジネスなどあらゆる状況で、様々な専門的知識だけでなく、社会とのかかわりから多様な価値観を含む知識基盤が必要とされている。ここで、基盤となる知識の量が IT 革新などの要因によって急速に増加しており、膨大な知識をユーザーが把握しきれなくなっていることを鑑みると、知識基盤から必要な知識を取り出して利用するには、それらの効率的な共有化を行うための方法開発を行う必要がある。本研究では大規模文書群の一例として第 1 期～第 3 期科学技術基本計画を用い、その中の知識が表出化された「キーワード」を定量的に分析する方法論を確立し、知識や情報の構造化・視覚化を分析するツールの開発を行う。さらにそのツールを政策や研究、ビジネスなどの支援に用いることの可能性、問題を考察する。

キーワード：テキストマイニング，ネットワーク分析，知識構造

1. はじめに

IT 革新を経て膨大な知識が絶え間なく生み出される中で、その知識は社会において適切に共有されることが難しく、工業社会から知識社会への移行が謳われる現在では膨大な知識の活用が大きな課題となっている。時代と共に新たな価値観が数多く生まれ、語句 1 つ取ってもそのコンテキストに多様性が含まれるというのがその理由の一つである。

そのような知識の適切な共有において、どのように知識体系が異なっているかを可視化・明示化することが有効であるが、このための方法論とツールは不足している。そこで本研究では、「文書群」内の語句のコンテキストをネットワーク分析で視覚化するツールを開発することで、異なる時代や組織でのコンテキストの違いを明示化し、語句に関する知識を共有化を促進しようと試みる。

またその際の分析対象として、ある特定問題領域に取り組む「知識コミュニティ」を扱う。ある専門的知識をストックしている知識コミュニティを対象知識の構造化・視覚化を行い、ユーザー側の専門的ではあるが限定的なニーズを満たすことで、社会で利用可能な知識基盤となるツールを開発するにあたり、知識基盤としてのツールの利用性を高めることにつながるからである。

2. 知識構造マップの開発

語句のコンテキストを視覚化するために開発した「知識構造マップ」は（1）テキストマイニングによる形態素解析、および語句の係り関係を表すクロス表作成（2）クロス表の正方行列化（3）作成した正方行列のネットワーク分析、という作成手順を踏んで作成される。以上の手順において、本研究ではテキストマイニングツールとして Text Mining for Clementine10.1（以下 Clementine）、ネットワーク分析ツールとして UCINET 6 for Windows（以下 UCINET）を用いた。

まず大規模文書群の電子テキストデータをテキストマイニングによって形態素解析し、ノイズ処理をした上で Clementine のクロス表ノードを用いて主要語（今回は名詞のみ抽出）と係り先主要語（主要語に係る名詞）のクロス表を作成する。ただし、この時点でのクロス表は正方行列でないため、以後必要なネットワーク分析の計算ができない。そのため、クロス表を Excel 上に移し、VBA で正方行列に変換するためのプログラム処理を行う。

次に、作成した正方行列を UCINET に保存し、同ツール内に含まれる NetDraw で正方行列を読み込み、有向グラフによるネットワークを作成する。さらに正方行列から自動計算された中心性指標をノードに反映させることで、語句の重要性、関係性を視覚化する。このような過程を経て得られたネットワーク図を「知識構造マップ」と呼ぶことにする（図 1）。

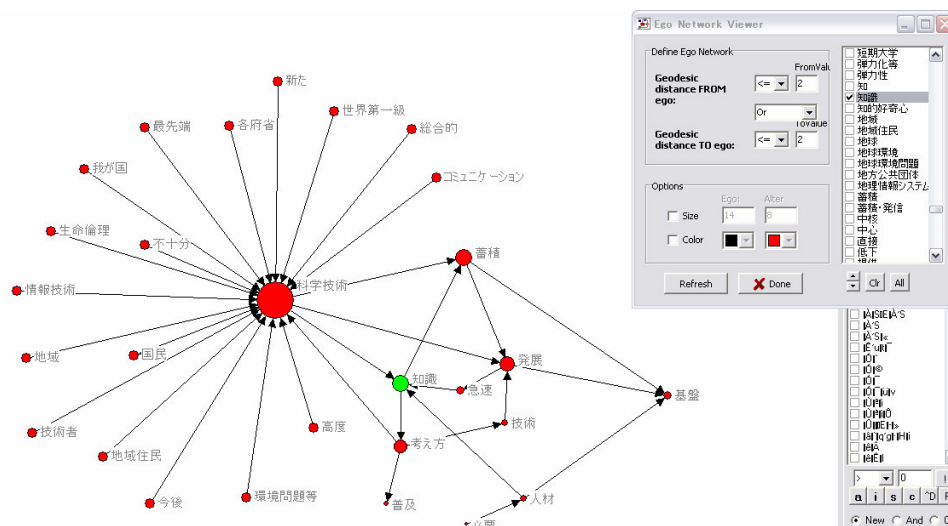


図 1. 検索語句「知識」を中心とした知識構造マップ（距離 2 のエゴセントリックネットワーク）

3. 事例分析

3. 1 分析対象

本研究では、特定の知識領域を持つ組織（知識コミュニティ）として研究・技術計画学会（JSSPRM）を対象とした。研究・技術計画学会は、技術経営の向上、科学技術関連政策の立案と推進など、科学技術の経営・政策全般にわたる研究交流と情報交換を図ることを目的としているため、我が国においてイノベーション政策に関連する諸概念が論考として集積された場であるといえる。今回研究対象とした文書群は、1986 年～2007 年度に学会へ提出された研究・技術計画学会の予稿など 1917 本である。

3. 2 分析結果と考察

分析結果として「研究開発」を検索語句例とした知識構造マップを紹介する。対象文書群の 1986 年から 2007 年までの「研究開発」の周辺概念変化を知るために、本研究では対象文書群を 4 期間に分けて知識構造マップを各々作成することにした(図 2)。

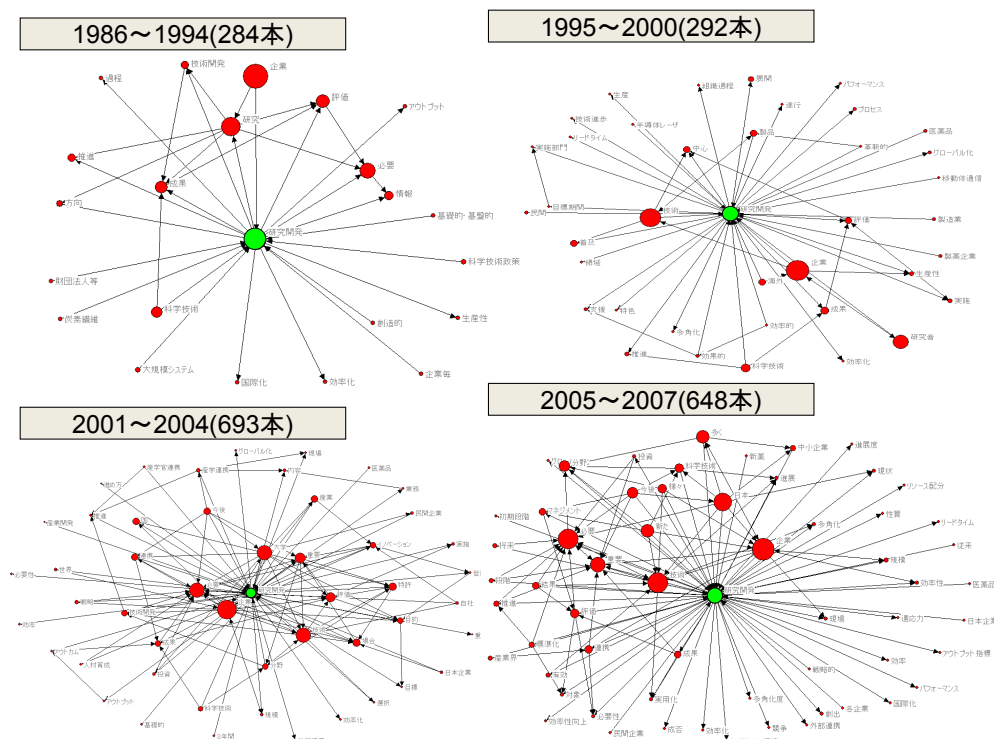


図 2. 「研究開発」を検索語句とした知識構造マップの時系列変化

そして時系列で作成した知識構造マップを基に、研究・技術計画学会における「研究開発」周辺の変化について考察を行う。なお、分析の都合上、知識構造マップに載せたノードは文書群内に3回以上リンク（語句のつながり）が出現したノードのみを抽出している。

＜研究開発の種類に関する時系列変化＞

第1期で「基礎的・基盤的」という語がつながっているのに対し、第2期では「医薬品」、「半導体レーザ」、「移動体通信」といった研究開発の具体的内容がつながっている。「医薬品」に関してその傾向は第3期、第4期でも変わらず、研究開発の種類として「医薬品」が大きなトピックになっていると考えられる。

これらの時系列的变化から、研究・技術計画学会では初期に基礎・基盤研究を、そしてその後精密機械や医薬品の研究開発にトピックの焦点が絞られていくといった傾向があったと考えられる。

＜研究開発の性質に関する時系列変化＞

第1期～第4期を通して「効率化」という語句が見られ、一貫して研究開発の効率性を求める記述があることが分かる。また、第1期で「創造的」、第2期で「革新的」、第3期で「イノベーション」、第4期で「創出」という語が周辺にあることから、言い回しは変化するものの、研究開発において常に新たな創造が求められていることが見て取れる。

＜研究開発の体制に関する時系列変化＞

第1期～第4期を通して「企業」は大きく関わっているが、第1期では「財団法人」、第2期～第4期では「民間」と、企業の質は変化していると感ぜられる。また、第1期では「企業毎」の研究開発だったのが、第3期からは「産学官連携」、「外部連携」の研究開発となっていることから、研究開発の組織的枠組みを変える動きが見られる。さらに第2期から第4期にかけて「海外」、「グローバル化」という語が出現していることから、研究開発の地理的範囲も拡大していると考えられる。

＜研究開発のアウトプットに関する時系列変化＞

第1期から「アウトプット」、「成果」といった語が研究開発の周辺語句として見られるが、特に第4期において「進展度」、「多角化度」、「アウトプット指標」のような量的にアウトプットを示そうとする語句が出てきていることから、近年では特に研究開発成果の定量化を試みる動きがあることが分かる。

4. まとめ

4. 1 結論

1986年から2007年までの研究・技術計画学会の論文集を事例にして、「研究開発」を中心とした知識構造マップを4期に分けて作成し、それらの時系列比較から「研究開発」とリンクする複数の周辺概念に関する変化を分析した。知識構造マップによって視覚化された文書群内語句のネットワークは、注目した語句の周辺概念について種類や中心性を明示しており、その結果より、大規模文書群の中から注目した概念の使われ方を抽出し、それらを共有するにあたってのツールの有用性を確認できた。

4. 2 今後の展望

本研究の利用性を高めるために、想定可能な具体的ユーザーを交えたワークショップを開催し、そこで知識構造マップの利用可能性についての議論を行ったところ、以下のようなコメントが得られた。

[判例データベースに適用した場合]

・裁判員制度が実施され、法律に疎い市民が担当する裁判に関連した知識を必要とする場合を考えたとき、知識構造マップではこれまでのような語句検索ではなく、語句の結びつきをみることができるので、「大学」と「パワハラ」が結びついていれば大学関連でパワハラ訴訟が多い、というような推測が可能となる。従来の条件検索よりもユーザー側に気づきを与えることが出来るだろう。

[特許データベースに適用した場合]

・現在ニーズとしてあるのは、特許出願の先行技術の調査をしたいという場合に、通常はこれから出そうとする特許データの全文を入力して、それと類似のものを検索したいというものである。その一つの

アプローチとして知識構造マップは役に立つと考えられる。

[論文データベースに適用した場合]

・科学技術政策の立案においては、例えば電子情報通信分野だったら、その分野の研究開発投資分野を洗い出すときに、最初に俯瞰をして、重要テーマについてシナリオ作りを行う。論文データベースから知識構造マップを作成できれば、エキスパートのレビューと照らし合わせながら **Evidence** に基づく議論が政策担当者側とエキスパートの間で可能になる。政策の現場では可視化ツールを常に必要としているため、実際のニーズはかなりあると思われる。

これらの議論から、知識構造マップのような可視化ツールを政策や研究、ビジネスなどの支援に用いる可能性は大いにあると思われる。ただし、結果からどういう示唆を与えれば上記のようなニーズにこたえる形になるのかという方法論が学術的コミュニティでもきちんと議論されていないため、ツールの技術をもとにどのような全体の仕組みづくりを行っていくかという点が、今後の課題になると考える。

<参考文献>

大隅昇「データ型のテキストマイニングー最近の動向とそれが目指すものー」テキストマイニング研究会，2006.

金光惇「社会ネットワーク分析の基礎 社会的関係資本論に向けて」勁草書房，2003

安田雪「実践ネットワーク分析」新曜社，2001

片上大輔，清水英明，田中貴紘，新田克己，山田隆志「文献情報に基づく学際的分野間ネットワーク分析」，人工知能学会全国大会論文集 1B2-07 ， 2007.