テキストマイニングによるインドネシアにおける水資源インフラストラクチャニーズの分析

独立行政法人水資源機構　杉浦政裕*1

近年、日本国内では積極的に水関連事業の国際展開が検討されている。こうした状況の中で、絶えず変化する国際市場におけるニーズ分析を行うことは、日本の水関連事業の国際展開を考えるうえで不可欠である。

幸いにも私たちは、情報通信技術の普及により膨大な情報にアクセスしやすい環境にある。一方で、発展途上国では、情報そのものが時系列に整理できるほど整えられていない場合もある。いずれにしても、求める情報を提示すれば、自動的に膨大な情報からニーズを分析し、提供してくれるようなシステムは現段階では存在しない。人工知能が発達してきたとはいえ未だ創造力思考を行っているのは人間である。そのため、分析者のもつ知識、事業経験、現場感覚などを引き出す「気づき」を支援することが求められる。

そこで本研究では、情報そのものが時系列に整理できるほど整えられていない場合の水資源インフラの整備・管理事業のニーズ分析手法を提案する。提案する分析手法は、単年の新聞記事から的確な情報（頻出する水資源に関連する単語）を取り出し、関連する単語の位置関係を可視化することにより、分析者の「気づき」を支援する方法である。そして本研究では、提案する方法の適用限界を示すとともに、現在あるいは将来的にも日本経済にとって重要となるインドネシアの水資源インフラの整備・管理事業のニーズ解明・把握を試みる。

【キーワード】インドネシア、事業計画、数量化Ⅲ類、テキストマイニング、ニーズ分析

1．はじめに

近年、国内外の水問題解決を目指した「チーム水・日本」や水ビジネスの国際展開など積極的に水関連事業の国際展開が検討されている。こうした状況の中で、絶えず変化する国際市場におけるニーズ分析を行うことは、日本の水関連事業の国際展開を考えるうえで不可欠である。特に、対象国全体の社会状況を大まかに把握し、社会ニーズの所在とその位置づけを知ることは重要である。しかし、対象国全体の社会状況を大まかに把握し、社会ニーズの所在とその位置づけを知ることは、容易なことではない。

幸いにも私たちは、情報通信技術の普及により膨大な情報にアクセスしやすい環境にある。膨大な情報は、情報が大量かつ多様という二面性をもっている。この二面性は、利用者にとって大変やっかいな性質である。利用者は求める情報を得るためには、文書の内容を精査しなければならないため、その煩雑さも膨大になるからである。一方で、発展途上国では、情報そのものが時系列に整理できるほど整えられていない場合もある。いずれにしても、求める情報を提示すれば、自動的に膨大な情報から精査して提供してくれるようなシステムは現段階では存在しない。人工知能が発達してきたとはいえ未だ創造力思考を行っているのは人間である。その創造力思考を支援するひとつの方法は、「気づき」を支援することであろう。

*1　本社人事部 048-600-6500

2. 研究の課題と方法

(1)研究の課題

本研究の課題は、次の２つである。ひとつの課題は、現在あるいは将来的にも日本経済にとって重要なアジア地域、中でも日本の建設産業にとって有望市場であり、日本と同じ稲作文化圏でもあるモンスーンアジア地域に位置するインドネシアの水資源インフラの整備・管理事業のニーズ解明・把握を試みることである。もうひとつの課題は、収集情報に制約がある場合の分析手法を検討するために、単年分の新聞記事によるニーズ分析の結果とその限界について考察することである。

(2)研究の方法

本研究においては、杉浦(2010) により提案された複数年の新聞記事をテキストマイニングして社会状況をマッピングすることによる時系列的に社会状況の変化を把握する手法の研究と同様に、社会の中に潜む水資源インフラ整備ニーズに焦点を合わせるため、客観性が高く収集しやすい新聞記事の分析を採択する。新聞記事を代用とする理由は、①事実を客観的に伝えることにより発達してきたマスメディアの中でも特に影響力を持つ、②マスメディアを通じて膨大な情報が流通しており、マスメディアの動向を知ることは社会ニーズを考えるうえで重要となると判断したからである。

新聞は、インドネシア国内で発行されている現地新聞『じゃかるた新聞』を選定する。分析対象は、水資源インフラ関連記事（じゃかるた新聞社のオンラインデータベースに蓄積されている 1 年分の記事）とする。

杉浦(2010)の研究では、日本の主要な新聞である『毎日新聞』における水資源インフラ関連記事（1992 年、1996 年、2000 年）を用いて、社会状況の変遷をマッピングし分析することにより、社会状況変化を先取りしたインフラ整備計画策定のためのニーズの芽の発見支援の手法を確認しているが、ここでは社会状況の時系列的な変化に関する分析が難しいとされる１年分の新聞記事によるニーズ分析の結果とその限界について考察する。

3. 先行研究の概説

テキストマイニングによる分析手法の研究は、記事の動向表現の分析、記事の因果関係の分析、潜在的意味解析、グラフ理論の応用、ニューラルネットワーク理論の応用などが行われている。また、実務者向けの簡易な分析手法の研究も行われている。

まず、テキストマイニングにより記事の動向表現を分析する関連研究は、数詞に注目してその周辺の言語パターンを解析することにより情報を分析する方法（斉藤ら（1998））、係り受け関係を利用する方法（藤畑ら(2001)）、統計量名を注釈付けするためのタグセットを定義してアノテーション付コーパスにより機械学習を使って自動抽出する方法（森（2007））、統計量表現に共通してよく出現する suffix に着目したパタンマッチングを利用した統計量表現抽出する方法（河合ら（2008））、そして、複数の記事に時系列に出現する様々な動向情報からデータを取り出し、データテーブルを作成し、Data Transaction, Visual Mapping, View Transformation の３つのプロセスからデータを可視化表現する方法（松下ら（2005））などがある。

次に、テキストマイニングにより記事の因果関係を分析する関連研究は、諸事象間の因果関係を有向グラフとして表し、事象の連鎖反応を分析する手法（佐藤ら(1999)、佐藤ら(2006)）、因果関係を含む可能性の高い共起関係に着目し、共起ネットワークを構築・観測する方法（河合ら（2008））、そして、因果関係知識の自動獲得を目指して、任意に定めたテキスト集合に対して因果関係情報に注釈をつけることにより、因果関係の出現傾向を分析する手法（乾ら（2005））などがある。

記事の動向表現の分析や記事の因果関係の分析は、高出現頻度語の分析を中心とし、既知の傾向の定量的な把握には貢献している。しかし、これらの手法では、同様の意味をもつ低出現頻度重要語を取り込むことは困難であった。そこで、語句の背後にある意味を分析することにより、低出現頻出重要語を分析の対象に取り込むことを試みている手法が、テキストの潜在的意味解析（LSA:Latent Sematic Analysis）である。テキストの LSA の関連研究は、特異値分解に基づいた LSA を発展させた PLSI（Probabilisitic Latent Sematic Indexing）、因子分析と情報理論に基づいた堅固な統計モデルによ

る SLSI(Statiatical Latent Semantic Indexing)などがある。LSA は、中村（2008）が解説するとおり、すべての文章の背後には意味の構造が存在すると考え、これを行列の形で表現し、分析するところに特徴がある。また、LSA は、さまざまな言葉で表現される意味の豊かすぎる部分を、行列の分解という形でとり除き、複数の語句の背後に共通して潜在する意味構造を抽出している。それは、漠然とした意味の豊かさよりも、凝縮した構造の方が語句に留まらない意味の豊かさを効率的に表現できる可能性があるからである。

さらに、グラフ理論やニューラルネットワーク理論を応用することにより、既存文書の分析から未来予測を試みる研究も取り組まれている。

実務者向けの簡易は分析手法の研究は、膨大な文字情報の中から高出現頻度語を抽出し、それらを数量化Ⅲ類により、抽出された各語の相対的位置関係をマッピングすることにより社会状況の可視化を図り、実務者が社会ニーズに「気づく」ことを支援する方法（杉浦(2010)）がある。杉浦(2010)の方法は、実務者が公共事業のニーズ分析を行う手法であるため、手軽に分析でき、実務者の事業計画の経験や現場感覚を十分に引き出すための「気づき」を支援することに重点を置いている。

4．テキストマイニングによる地域ニーズの分析

(1) 分析対象データと分析手順

a）分析対象データの準備

①『じゃかるた新聞』全文電子データ（2008 年 8 月 1 日から 2009 年 7 月 31 日まで）を使って、検索キーワード（水資源、洪水、渇水、水道、水質、水力、用水、地下水）で各年の記事見出しを検索し、記事を抽出する。以下、抽出された記事を分析対象とする。

②頻出単語から、関係が深いと考えられる単語を抽出し、それを変数として選択する。

③置換辞書を作成するために、分析対象を分かち書きし、名詞を抽出する。

④抽出された名詞から、類似の意味をもつ名詞の表現を統一するために、置換辞書を作成する。「ソロ川」＝「河川ー川系」、「チリウン川＝河川ー川系」と表示する。

⑤作成した置換辞書を使って、分析対象の類似の意味をもつ名詞表現を統一する。

なお、分かち書き処理、キーワード抽出、置換辞書作成には『Word Miner』（日本電子計算機社製）、数量化Ⅲ類分析には『エクセル統計 2006』（社会情報サービス社製）を使用した。

b） 単語の出現頻度の順位

①作成した置換辞書を用いて、再度、全分析対象を分かち書きし、記事毎に名詞を抽出する。

②抽出された名詞毎の出現数を数える。なお、対象記事の長さによる同一出現語の出現頻度のバイアスを除去するために、対象記事 1 つに複数回出現する同一語は、対象記事 1 つにつき 1 回の出現とした。

③高出現頻度語 100 を目安に、出現頻度数の多い順に並べる。

この手順により、社会問題のキーワードの傾向をとらえる。

c） 数量化Ⅲ類による高出現頻度語の相対的位置関係のマッピング

①対象記事中に、「4.(1) b)③」で並べられた名詞の有無を調べ、その結果を行列データにする。

②行列データを元に、数量化Ⅲ類分析をする。この手順により、新聞記事から社会状況をマッピングし、社会ニーズの状況を定量的にとらえる。

③分析にあたっては、単相関係数 r が 0.5 以上または累積寄与率 50%以上となることを目安に分析する。一般に、数量化Ⅲ類においては、単相関係数 r や累積寄与率の値が大きいほど情報が集約された結果が示されるが、これらの値の大小に関する統計学的な基準は存在しない。ここでは単相関係数 r が 0.5 以上または累積寄与率 50%以上となることを分析の目安としたが、単相関係数 r が 0.5 未満または累積寄与率 50%未満であっても分析結果は必ずしも棄却されるわけではない。

(2) 単年度分析結果とその限界

a） 分析対象データのプロフィール

表４－１　キーワードにより選別された関連新聞記事数

キーワード	キーワードにより選別された関連新聞記事数 2008年８月１日から 2009年７月31日まで
水資源	10
洪水	89
渇水	1
水道	27
水質	10
水力	12
ダム	24
用水	3
地下水	3
計	189
計（重複記事除く）	177

（筆者作成）

『じゃかるた新聞』全文電子データ（2008 年８月１日から 2009 年７月 31 日まで）を使って、検索キーワード（水資源、洪水、渇水、水道、水質、水力、用水、地下水）で１年間の記事内容を検索し、抽出された記事を表４－1 に示す。分析対象記事数は重復記事を除く 177 件であった。

分析対象のキーワード数を、表４－2 に示す。置換辞書使用後の全文を通じて異なるキーワード数は、1,732 個であった。置換辞書使用後の全文を通じて異なるキーワード数は、①高出現頻度語から関係が深いと考えられる単語を抽出、②置換辞書を作成し類似の意味をもつ名詞を統一表現にする、③全文を通じて異なる名詞の数を計数したものである。

今回は 220 語を置換辞書、記号・句読点・助詞を削除辞書として登録し、置換辞書および削除辞書により、ジャカルタ市や東ジャワ州など地方自治体名は「自治体ー州県市」、チタルム川やソロ川などの河川名は「河川」、危機や非常時など危険や非常を表す表現は「危険ー非常系」などへ表現を統一し、記号・句読点・助詞は削除した。

b)　単語の出現頻度の順位表示および数量化Ⅲ類分析によるインドネシア国内発行新聞（2008/2009 年）のテキストマイニングによる単年分析の結果

2008/2009 年（2008 年８月１日から 2009 年７月 31 日まで）の高出現頻度単語を単年分析し、出現頻度順に整理した結果を表４－３に示す。また、表４－３に表示されている高出現頻度語を記事毎に単語

表４－２　分析対象キーワード数

	2008/2009
置換辞書使用前の全文を通して異なるキーワード数	4,539
置換辞書使用後の全文を通して異なるキーワード数	1,732

（筆者作成）

の有無を行列データにし、数量化Ⅲ類分析をした結果を図４－１に示す。

数量化Ⅲ類では、軸の解釈に関連する計算はしていないため、軸の解釈にはこだわらない。そのため、図４－１の解釈にあたっては、単語の相対的な位置関係のみに注視する。その結果、分析者の経験や能力によりさまざまな解釈の仕方が誘発されるであろう。複数の分析者により図４－１を分析した場合、議論の幅が広がり、各分析者の更なる気づきを期待できる。

なお、分析にあたっては、単相関係数 r が 0.5 以上または累積寄与率 50%以上となることを目安にしたため、行列データの行や列に１や０が多い場合は、その行や列を削除して、行列を作成し直した。そのため、表４－３の高出現頻度語は、図４－１に表示されている語数よりも多い。

図４－１は、たとえば次のように解釈する。語群⑤は 2008 年３月下旬に発生したギントゥンダムの決壊により、百人以上の死者を出した事故であったが、一過性のトピックであるため、全体の語群から離れたところに位置している。しかし、中央政府による貯水池施設の安全点検、メディアによる災害時の避難体制への高まりがあることから、語群⑥、語群④へ派生していることがわかる。語群⑥では、事故原因の追究を行い、今後更にダム安全性を向上させ、国民が安全・安心して生活できるよう努力していることを示している。また、語群④では、メディアは、死者数、事故原因の報道を通じ、避難体制の整備やダム安全性の確保の必要性を社会に投げかけていることを示している。

語群①は中央政府のＯＤＡ依存状況を示しており、ＯＤＡにより国土インフラの整備を推進しようとしている。特に、水道事業への関心は大きい。

また、語群②は中央政府は、同時期に計画的な住

表４－３　高出現頻度語の順位表示

2008/2009年(2008.8.1-2009.7.31)

	関連頻出名詞	サンプル度数
1	自治体-都道府県市系	144
2	地域-地域系	140
3	洪水系	100
4	降雨-雨系	75
5	事業-工事系	66
6	全国-インドネシア系	66
7	被害	66
8	発生	65
9	市民-住民系	54
10	危険-非常系	53
11	計画-予定系	53
12	対策-対策系	51
13	政府	45
14	河川-川系	42
15	環境	35
16	中央-中心系	35
17	影響	34
18	公共	34
19	指摘	34
20	水	34
21	水道水-水道系	34
22	貯水-貯水系	34
23	原因-理由系	33
24	支援	31
25	必要	31
26	開発	30
27	方針	30
28	管理	29
29	死亡	28
30	道路	28
31	周辺	27
32	協力	25
33	生活	25
34	決壊	24
35	整備	24
36	説明	24
37	知事	24
38	地元紙	24
39	改善-改善系	23
40	使用	23
41	施設-設備系	23
42	事故	23
43	避難	23
44	活動	22
45	局長	22
46	調査	22
47	発表	22
48	確認	20
49	予算	20
50	ギントゥン	19

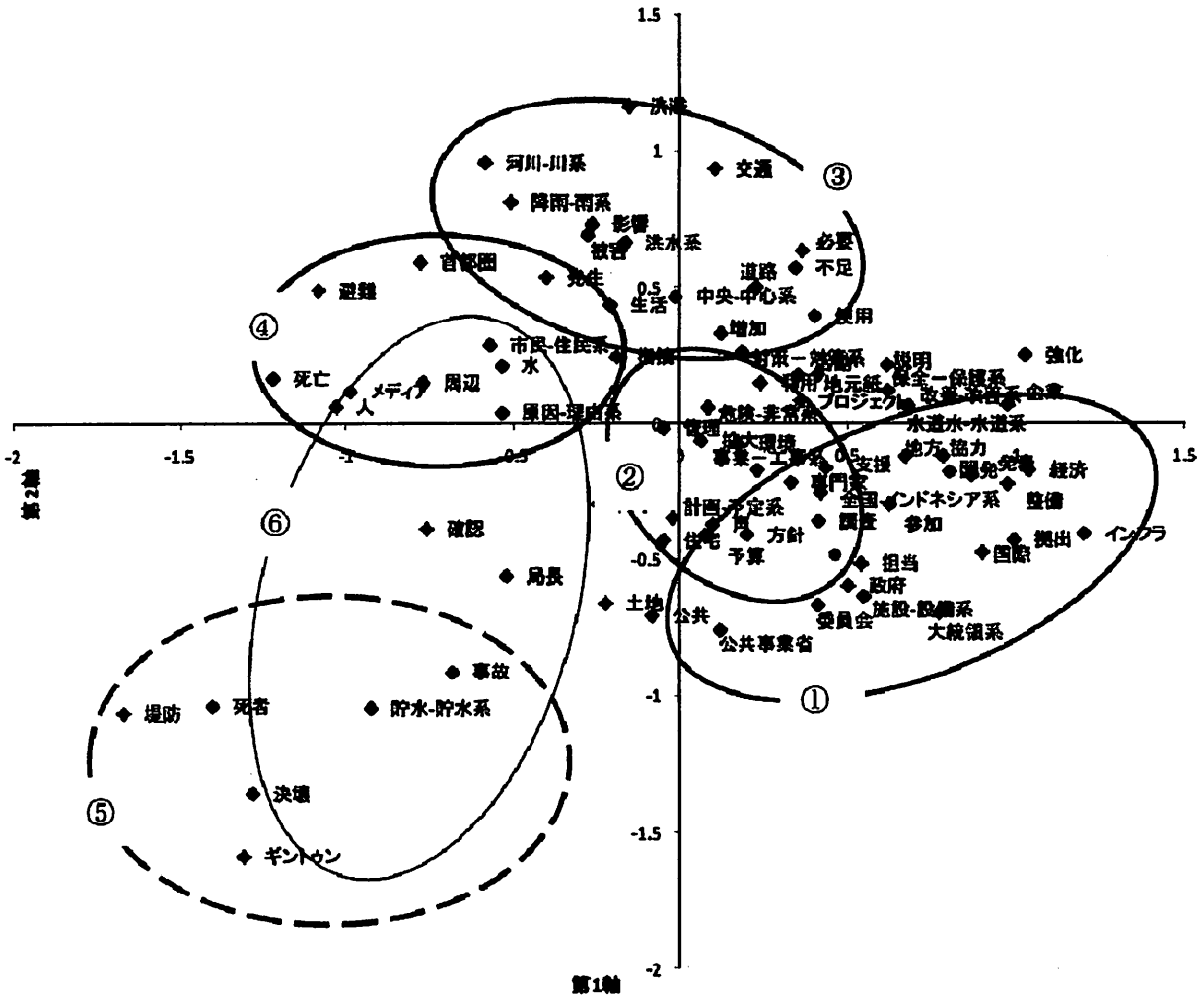
	関連頻出名詞	サンプル度数
51	プロジェクト	19
52	会社	19
53	拡大	19
54	首都圏	19
55	企業	18
56	参加	18
57	死者	18
58	メディア	17
59	悪化	17
60	国内	17
61	声	17
62	利用	17
63	交通	16
64	公共事業省	16
65	国際	16
66	市内	16
67	住宅	16
68	設置	16
69	専門家	16
70	土地	16
71	委員会	15
72	監視	15
73	機関	15
74	強化	15
75	経済	15
76	国営	15
77	渋滞	15
78	大規模	15
79	大統領系	15
80	中部	15
81	不足	15
82	保全-保護系	15
83	民家	15
84	インフラ	14
85	ダム	14
86	提出	14
87	災害	14
88	実施	14
89	人	14
90	増加	14
91	担当	14
92	地方	14
93	堤防	14

（筆者作成）

宅整備を図っていることもわかる。

語群③は降雨による洪水被害は、都市部の道路が冠水することが多く、生活の支障をきたしていることがわかる。

これらの解釈は、単年分の水資源インフラに関する新聞記事から出現頻度の高い単語をキーワードとして抽出し、個々のキーワードの関連性を一つの図上に集約・可視化することにより、分析者に対して



（筆者作成）

図４－１　数量化Ⅲ類分析による2008/2009年の社会状況のマッピング

ストーリーを伴うニーズへの「気づき」が促され、可能となった。

これらの状況から、流域水ユーザー、水関連インフラ整備のために必要な技術的条件・法制度の視点から、次のようにニーズを捉えることもできるであろう。新規事業（水道施設建設など）、改築事業（道路改築、ダム施設点検・補修など）などのインフラ整備事業が望まれている。それと並行して、ダムなど河川構造物の安全基準の見直し、警戒避難のためのハザードマップ作成、自治体への災害復旧支援制度、市民への生活再建支援制度などの技術基準や法制度の策定ニーズがあると解釈することができる。

この手法は、複数年の社会状況の可視化と比較して、得られる情報量が単年分に限られるため、分析者へ与える「気づき」の支援は小さくなっている。しかし、社会的問題、社会的問題の関係、社会的問題解決のための方策など多くの知見に気づくことができることを示している。

c) 単年度分析の限界

分析結果から、単年度分析の限界と活用の可能性ついて以下に述べる。

単語の出現頻度の順位表示および数量化Ⅲ類分析によるインドネシア国内発行新聞（2008/2009年）のテキストマイニングによる単年分析では、社会問題のキーワードとして長期化する単語、短期間でキーワードとして存在しなくなる単語、出現頻度が増加している単語の分析できない。

これは、単語の出現期間の面と単語の出現頻度の面から、以下の分析の限界を示している。単語の出現期間の面からは、基本語となっており社会の変化を表現していない単語と社会の変化を表している可能性が高い単語による区別を困難にしている。そして、単語の出現頻度の面からは、出現頻度が急増・急減する単語から社会変化の発見を困難にしている。

上述の限界を全て克服しなければ分析できないということはない。単年のみの記事であるため先の予測は容易ではないが、過去の判例や行政の判断パターンを学ぶことにより、単年の少ない手がかかりから、確からしさを高めた「気づき」を支援することが可能となる。

5. まとめ

本論文では、単年分の現地新聞記事により、インドネシアの水資源インフラの整備・管理事業のニーズ解明・把握を行った。併せて、単年分の新聞記事によるニーズ分析の限界について言及した。これらにより、多変量解析の限界である低出現頻度の重要単語の発見や単年度分析の限界である時系列変化を把握することの困難性を認識しておくことが必要であるが、社会的問題、社会的問題の関係、社会的問題解決のための方策など多くの知見に気づくことができることが明らかになった。

本手法（テキストマイニングによるキーワード抽出と数量化Ⅲ類によるキーワードの関連性の可視化）を利用することにより、情報そのものが時系列に整理できるほど整えられていない場合であっても、分析者の知識と経験を引き出す「大きな気づき」に繋がる支援を期待できることが示された。

また、例えばインフラ整備供給者側の本手法の利用方法としては、文字化した住民と行政組織との対話、インフラ整備に関する会議の議事録、文章で書かれた住民アンケート結果などのデータを分析することにより、分析の幅を広げることができ、分析者の「気づき」をさらに支援することが期待できる。これは、インフラ整備供給者側のみならずインフラ整備事業の受注者側にとっても、将来のニーズを先取りする有効な手法となるであろう。

今後さらに水資源インフラに関するニーズの分析やニーズの予兆発見に接近するためには、多変量解析の限界を克服しなければならない。特に、意味の共起関係を探るための単語ネットワーク分析の研究は重要となる。並行して、分析に不要なノイズの切り捨て、重要な単語を抽出し分析目的に合致した意味や評価を行うために情報源の加工作業である「辞書」の整備を進めることは不可欠である。これらの研究が進展することにより、テキストマイニングの応用は益々広がり、日本国内外の公共事業ニーズやシーズ分析支援が盛んになっていくと考える。（本研究の一部は、平成20年度（財）日本建設情報総合センターの研究助成を受けて実施したものです。）

註

中村健太郎（2008）「潜在的意味解析」豊田秀樹監修『データマイニング入門－R で学ぶ最新データ解析』東京図書，pp. 271-272.

参考文献

Makoto Suzuki(2008)Text Categorization Using the Maximum Ratio of Term Frequency, *J Jpn Ind Manage Assoc*, 58, pp. 438-444.

石田基広(2008)『Rによるテキストマイニング入門』森北出版.

市村由美，長谷川隆明，渡部勇，佐藤光弘(2001)「テキストマイニング－事例紹介」『人工知能学会誌』人工知能学会，No. 16，Vo. 2，pp. 192-199.

乾孝司，奥村学(2005)「文書内に現れる因果関係に出現特性調査」，『情報処理学会研究報告』，情報処理学会，pp. 81-88.

大澤幸生(2003)『チャンス発見の情報技術』東京電機大学出版局.

大澤幸生(2006)『チャンス発見のデータ分析』東京電機

大学出版局.
河合英紀，齋藤悠，土田正明，水口弘紀，國枝和雄，山田敬嗣（2008）「新聞記事における統計量表現の共起ネットワーク」『第22回人工知能学会全国大会論文集』人工知能学会，3K3-10.
川前徳章、青木輝勝、安田浩（2002）「統計的潜在的意味空間の抽出」『自然言語処理(148-4)』、情報処理学会，pp. 25-30.
斉藤公一，迫田昭人，中江富人，岩井禎広，田村直良，中川裕志（1998）「数値情報をキーとした新聞記事からの情報抽出」『情報処理学会研究報告(98-NL-125)』情報処理学会，vol. 98、No. 48，pp. 63-69.
佐藤浩史，笠原要，松澤和光(1999)「テキスト上の表層的因果知識の獲得とその応用」『電子情報通信学会技術研究報告（TL98-23）』電子情報通信学会，pp. 27-34.
佐藤岳文，堀田昌英(2006)「Web マイニングを用いた因果ネットワークの自動構築手法の開発」『 社会技術研究論文集』社会技術研究会，Vol. 4，pp. 66-74.
白井康之，小関悠，小池亜弥(2009)「テキストマイニングによるトレンド情報抽出環境の構築技術」『三菱総合研究所所報』三菱総合研究所，No. 51，pp. 110-123.
新納浩幸(2007)『R で学ぶクラスター分析』オーム社.
杉浦政裕(2010)「テキストマイニングによる簡易な社会状況可視化手法の開発－新聞記事を材料として－」『経済科学論究』埼玉大学経済学会，No. 7，pp. 99-110.
高橋由光（2005）「ネットワーク分析を利用したテキスト・マイニング-2004 年の朝日新聞を事例にして-」『2005 年人文科学とコンピュータシンポジウム』情報処理学会，pp. 35-40.
那須川哲哉(2006)『テキストマイニングを使う技術/作る技術』東京電機大学出版局.
野村総合研究所(2008)『顧客の声分析・活用術』リックテレコム.
田村正紀(2006)『リサーチデザイン』白桃書房.
藤畑勝之，志賀正裕，森辰則（2001）「係り受けの制約と優先規則に基づく数量表現抽出」『情報処理学会研究報告（2001-FI-64，2001-NL-145）』情報処理学会，pp. 119-125.
松下光範、加藤恒昭（2005）「動向情報に基づく情報可視化の基礎検討」『第 19 回人工知能学会全国大会論文集』，人工知能学会，IE3-03.
森辰則，藤岡篤史，村田一郎(2007)「動向情報編纂にためのテキストからの統計量の自動抽出」『第 21 回人工知能学会論文集』人工知能学会，3H9-4.
和多太樹，関隆宏，田中省作，廣川左千男(2005)「単語の出現頻度に着目した病院評判情報の分析」『研究報告－音声言語情報処理（SLP）』情報処理学会，vol. 2005， No. 50，pp. 15-20.
以上。

Study on Needs Analysis for Water Resources Infrastructure of Indonesia by Text Mining

Masahiro SUGIURA

In recent years, Japanese government is actively pursuing discussion on the international water related projects. For example, through coming together from all sectors of Japan (politicians, government, business, academia, and civil society), "Team Water Japan" has the will to support water security activities around the world. In this situation, it is necessary to find the needs of water related projects appearing in the international market timely.

Fortunately, we are easy to access enormous information by the spread of information and communication technology. On the other hand there is not enough information for the analysis in the developing country. At

all events it is important to draw the knowledge, project experience, sentiments of the social life, and so on from the needs analyst for needs analysis. It is to support "awareness" of the analyst. Because there is not the system to make the result of needs analysis automatically.

Therefore I will approach two problems in this study. The one is needs analysis for water resources infrastructure of Indonesia by text mining. Another one is to consider the limitation of the grasp the social situation by text mining using newspaper articles for one year.