ノート　**Note**

複雑な問題に対する複数の専門家の文書をテキストマイニングし比較する手法の開発

長 谷 川　豊[1]

Method of Text Mining for Multiple and Complex Documents which Experts Created

Yutaka Hasegawa

(Received 22 December 2014, revised 23 January 2015)

Synopsis: After the occurrence of the accidents of nuclear plants in Fukushima-dai'ich, many people told ''Scientific information about radiation is difficult to understand''. There may be differences in the expression of experts on their thoughts. Documents concerning to the radiation are collected and text mining to those documents are performed. We know that there exist their characteristic words. Collections are carried out for the sentence only, since sentences are available to obtain on the Web. Multiple documents are referenced in the most often manners. A chart representation is given from documents by text mining tool. The differences are appeared at the chart.

Key Words: Text Mining, Co-Occurrence Network, Self-Organizing Map, Venn Diagram

1. 諸　　言

福島の原発事故の発生以降，科学情報が分かり難いのではないかという評価が多く見られた[1–2]。その，原因として専門家の表現の相違にあると考えた。そこで，多くの文書を収集し，その文書をテキストマイニングし，特徴語の相違を検出することにした。そのために，まず，文書を多く収集することにした。入手の容易さから，Web 上の文書に限定した。そして，放射能に関係するキーワードで，多くの利用者に参照されている文書を手に入れるために，Google アラートを使用して，長期に渡り情報を蓄積し，その中で最も多く参照されている文書を複数検出した。それをテキストマイニングで加工しチャートを作成した。その，チャートを元に分析，および，プログラムによる自動の分析を行った。そのことに関して，この研究ノートで報告する。

2. 研究背景と目的

福島原発事故以降の放射能に関する情報は，一般市民によく伝わっていないとの雰囲気がある。何度か行われた国民意識調査の結果もそれを示している。しかし，Web 上には多くの専門家が発信する情報もある。つまり，専門家と非専門家である一般市民の間を ICT の技術を使って埋めることが求められているのではないだろうか。そのことをテーマにこの研究を始めた。

[1] 国士舘大学　政治経済学部　非常勤講師
有限会社スプライン IT& 教育研究室 室長

3. 研 究 方 法

要約的に述べると次のようになる。

- Google アラートにより，「放射能」に関連した Web 上の文書を収集する。
- その中で頻繁に参照されている専門家または専門組織の名称を抽出する。
- その専門家の論文を，テキストマイニングし，特徴語を抽出する。ツールは，KH Coder を使用する。
- 特徴語を元に，放射能に関する辞書を作成する。
- そして，その辞書を元に各専門家の特徴を数値化し，（以下で述べる tf*idf 値の算出で数値化する）それを使ってチャートを作成する。
  （チャートは，文書毎の特徴語の類似性を元に，特徴語同士の関連を表現するものと，文書に現れる特徴語のパターンを元に文書間の類似性を表すチャートを作成する。）
- それらの，数値およびチャートを元に文書の類似度即ち，差異を分析する。

3.1. 文書の収集

Google アラートを用いて，長期にわたり「放射線」に関して多くの人に読まれる文書を収集する。それらの情報をもとにテキストマイニングし，多くの文書に参照されている文書を特定する。また，それらの文書をさらにテキストマイニングすることで，特徴のある文書を抽出する。

⑴　Google アラートについて

Google アラートの提供する情報の根拠は，ページランク理論であり，ページランク（PageRank）は，ウェブページの重要度を決定するためのアルゴリズムであり，検索エンジンの Google において，検索語に対する適切な結果を得るために用いられている中心的な技術である。このアルゴリズムの発想は，引用に基づく学術論文の評価に似ている。ただし，学術論文との違いは，あらゆる種類の Web コンテンツとのリンクをチェックしていることと，常に新たな情報発信があったか，また，どの程度の頻度でその記事が検索されているかの情報が加味されている点である。そのことにより，Google は，任意のキーに対して Web の利用者にとって最も重要と思われる情報を提供することができる。つまり，Google アラートに表れた専門家（専門組織）は一般大衆の目線に近いものと言える。

3.2. テキストマイニングについて

⑴　テキストからなにが導き出せるか

日本語の文書は，読点により文という単位の区切りを表す。また，改行により文と段落という単位の区切りを表す。よって，一つの文書は複数の段落をもつことができ，一つの段落は複数の文を持つことができる。通常はテキストファイルの最後が文書の最後を表す。形態素解析で単語の抽出が出来る。形態素とはこれ以上分割できないテキストの単位で単語にあたる。単語には，量のデータとしては出現回数（通常は文書全体における単語の出現回数）と文書数（検出された文書の数）が測定できる。属性のデータとしては品詞を得ることができる。つまり，単語に関してはその文字列と品詞および番号という識別情報と文書群の中で出現回数と文書数というデータが得られる。

⑵　特徴語の抽出方法

それぞれの文書には何らかの意味が有る。一般的に，その意味と深い関わりのある単語は高い頻度で出現すると考えられる。その度合いを表すのに tf 値を用いる。その式を以下に示す。

tf(d,t)＝文書 d における語 t の出現数÷文書 d の総語数　・・d および t はそれぞれ識別番号

しかし，最も多く使用される単語はむしろ形骸化した特に意味の無い単語である場合もある。複数の文書を探索する場合，他の文書で出現数が少なく，少数の文書で出現数が高い方が，その文書の意味に深く関わると言える。その度合いを表すのに idf 値を用いる。

idf(t)＝log（全文書数÷語 t を含む文書数）　・・・t は識別番号

この，二つの数値を掛け算した値で，対象とする文書群の中のその単語の重要度を示す。それを，tf* idf 値という。よって，この tf*idf 値の高い単語を抽出することで，文書群の特徴語を抽出することができる。

⑶　共起ネットワークについて

一組の単語の共起性がどう測られるか示す。

二つの単語の出現数 x と出現数 y があったとして，どちらか一方が出現した回数を |x∪y|，両方が出現した回数を |x∩y| とする。

Jaccard 係数＝|x∩y|/|x∪y| で計算される。

KH Coder の共起ネットワークは，「文書×抽出語」表のデータを元に抽出語の Jaccard 係数を計算しその数値の大きなものから線で結合し図を作成する。その際に，tf*idf 値の高い抽出語を優先するが，予め，利用者が品詞や出現数，文書数で抽出語にフィルターを設定することができる。また，この Jaccard 係数を用いてクラスタリングすることもできる。つまり，共起ネットワークと同様の単語の関係性を用いてより繋がりの強い単語をクラスターとしてまとめることができる。このクラスターを共起ネットワークで表示した場合，そのネットワークの媒体の中心性のある結合を見出すことができる。媒体の中心性とは，複数の単語が一つの単語を介する事によって接続している状態を指し示す属性で，KH Coder では共起ネットワークのノードの色を媒体の中心性の強さによって色分けできるようになっている。以下の図から分かるようにこの媒体の中心性にある単語が各単語との関係性を表現しているといえる。

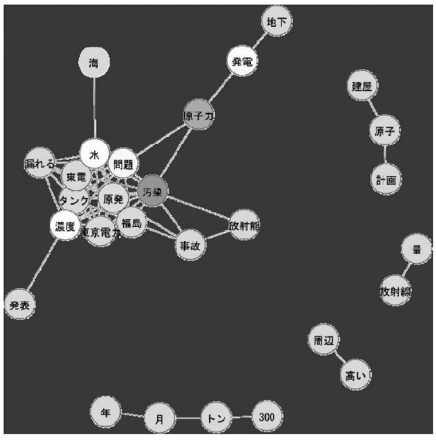

図-1　媒体の中心性のある共起ネットワーク

⑷　自己組織化マップ

自己組織化マップとは高次元のデータを，その特徴を踏まえつつ低次元で配置した図を作成する方法である。

今回はコホーネン（T. Kohonen）の自己組織化マップ（Self-Organizing Map）を使用した。複数データについて，複数個のパラメタのそれぞれのパターンの類似度を，多角的に重複して多数回学習する方式である。

4.　テキストマイニングの手続きと結果

4.1.　重要な記事の収集

Google アラートにより，17,577の文書を収集し，特徴語を抽出して，その中で専門家または，専門組織の名称を抽出した。

結果は，以下の様になった。

表-1　専門家または，専門組織の名称の出現数

専門家または，専門組織の名称	出現数
環境省	650
IAEA＝国際原子力機関	51
専門家（普通名詞および個人）	39
参議院	26
農林水産省	25
労働省	21
JAEA＝独立行政法人日本原子力研究開発機構	15
山形大学	15
東京大学	14
NSC＝アメリカ国家安全保障会議	13

トップ10以下は省略。以下には65の組織・団体が抽出された。これには，企業・マスメディアは含まれていない。この結果からそれらの専門家の文書をさらにテキストマイニングし以下のことが分かった。

⑴　多くの文書が参照している記事

- Google アラートおよび新聞のデータベースの放射線に関する重要な情報源を調査すると環境庁，原子力規制委員，IAEA＝国際原子力機関であると分かった。
- そのそれぞれの，情報の根拠をたどると「ICRP＝国際放射線防護委員会」の勧告にたどり着いた。
- また，さらにその勧告の重要な根拠となっているのが，放射線影響研究所の広島長崎の被爆者のデータである。

⑵　もっとも情報の多い文書

環境省の放射線健康管理担当参事官室の放射線に関する解説書が日本語では最も情報の多い文書である

ことが分かった。

その文書が「放射線による健康影響等に関する統一的な基礎資料 平成24年度版 ver. 2012001」である。

この文書を元に，リンクを辿ってパターンの異なる 7 つの文書を選択した。その内一つは新聞記事から特徴語を含んだ記事を探し出した。

⑶ 選出した 7 つの文書 （頭に付けた記号 C1～C7 を以降の文章で用いる。）

• C1 :「放射線による健康影響等に関する統一的な基礎資料 平成24年度版 ver. 2012001」

環境省：放射線健康管理担当参事官室

放射線の基礎知識と健康へ影響に関して，あらゆる情報を集めた文書。

• C2 :「放射性物質対策に関する不安の声について」

環境省

国民の不安にこたえ，不当な風評被害が生じることを避けるとともに，福島県内に住んでおられる方々の心情に鑑みた，環境省としての見解。

• C3 :「食品に含まれる放射性物質に関するガイドライン20130417-1」

厚生労働省

食品中の放射性物質への対応への文書

• C4 : 原爆被爆者の調査に基づくがんリスク推定値とその不確実性について

放射線影響研究所　日米共同研究機関

広島の原爆被害者の被爆状況を元とした世界に影響力のある研究施設の文書

• C5 :「原子放射線の影響に関する国連科学委員会の国連総会への2013年10月フクシマ報告書」

国連科学委員会

多くの専門家の良心の叫びのような報告書。

• C6 :「ICRP 勧告と基準値の考え方 ICRP 勧告と基準値の考え方」

ICRP：国際放射線防護委員会：専門家の立場から放射線防護に関する勧告を行う民間の国際学術組織

政府が安全性の基準としている文書

• C7 :「放射性物質，少量でも臓器に蓄積」

新聞記事

チェルノブイリ病理解剖を実施したベラルーシの医師の講演

4.2. 低線量の安全性に関する辞書の作成

辞書の構成として，「原因」，「結果」，「関係」，「対応」という構成要素を設定した。「低線量の放射線」に関する特徴語の中から，「原因」，「結果」，「関係」，「対応」に該当する単語を抽出した。「原因」には「放射線」，「結果」には「影響」，その間の「関係」を「被曝」，「対応」に「防御」を当てはめた。

このモデルに文書の特徴語を当てはめて，文書を分析した。この 4 つの単語以外の単語は，この 4 つの単語と無関係か，または，その言葉に関連するより詳細な実体に該当する単語か，または，これらの単語を含む大きな実体を表す単語に分類することができる。例えば，放射線には複数の種類がある。放射線とはそれらを統合した単語である。また，人体への影響も複数存在する。主に，低線量の被爆では癌があるが，その癌自体にも複数の癌がある。また，その原因と結果の間の現象である被爆にも多くのバリエー

ションがある。これらの単語をイメージとして，4 つの単語の配下に位置づけた。また，それらの単語の属性を表す単語も配下の単語とした。また，この辞書は名詞のみでできている。

そして，4 つの単語を統合した単語は上位の単語として，無関係な単語と同じ扱いとした。

例えば，社会や地域と言ったことは，上位のことであり放射線による汚染などの社会問題はこの辞書に含まれない。また，辞書に含まれる複数単語に関係する単語も辞書に含めることにした。それら関係をベン図で表すと以下のようになる。

4 つの主たる単語以外の単語については，すべて多くの文書に登場する重要な単語である。その重要度は，tf*idf 値によって確定されたものである。

7 つの文書を KH Coder で分析し，抽出された単語を該当する辞書のコードに登録する。そうして，それらの単語が出現したら該当するコードがカウントされるように論理を設定して，再び，7 つの文書を KH Coder にかけると，辞書のコードに関連した単語のみがカウントされる。また，その数値を元にチャートを作成すると，辞書に関連した文脈だけが表示される。

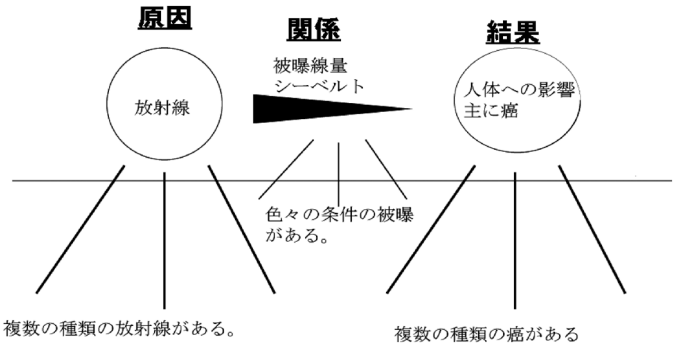

図-2 「原因」,「結果」,「関係」の図

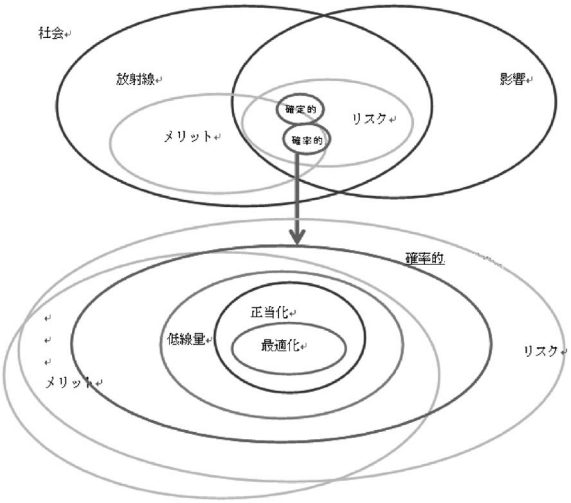

図-3 辞書に登録したコードのベン図

4.3. テキストマイニングの結果

段落単位で，左のコードが該当した段落が全体の何パーセントかを測定した。

表-2 7 つの文書における主要コードの出現パーセンテージ

コード	C1	C2	C3	C4	C5	C6	C7
身体的影響	4.87%	30.43%	0.00%	3.23%	0.84%	0.00%	33.33%
LNT モデル	1.01%	0.00%	0.00%	0.00%	0.00%	0.93%	0.00%
低線量	2.23%	0.00%	0.00%	17.74%	3.36%	3.74%	11.11%
被曝量	19.07%	39.13%	8.45%	35.48%	7.14%	16.82%	11.11%
放射線	11.76%	4.35%	0.00%	6.45%	0.84%	4.67%	0.00%
放射性物質	27.59%	17.39%	7.58%	4.84%	30.67%	4.67%	44.44%
疫学	1.42%	0.00%	0.00%	8.06%	1.26%	0.93%	0.00%
確率的影響	2.03%	8.70%	0.00%	0.00%	0.42%	0.00%	0.00%
影響	29.82%	34.78%	2.33%	20.97%	24.79%	8.41%	33.33%
被曝	29.82%	47.83%	0.58%	4.84%	20.17%	15.89%	33.33%
放射線被曝線量	19.07%	4.35%	4.66%	3.23%	2.52%	4.67%	0.00%
摂取量	5.68%	0.00%	7.00%	0.00%	1.26%	1.87%	11.11%
吸収線量	6.90%	8.70%	0.00%	8.06%	3.36%	0.00%	0.00%
被爆者	2.84%	0.00%	0.00%	1.61%	0.42%	0.00%	0.00%
リスク	38.54%	56.52%	2.33%	41.94%	22.27%	25.23%	33.33%
正当化	0.41%	0.00%	0.00%	0.00%	0.42%	0.00%	0.00%
防御	0.20%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
線量限度の適用	8.11%	0.00%	1.17%	17.74%	3.78%	4.67%	11.11%
防御の三原則	8.92%	0.00%	1.17%	17.74%	4.20%	4.67%	11.11%

※下線の引かれたコードは，全ての文書に現れている。この結果は，7 つの文書に関してコードに関連する低線量の放射線の安全性についての特徴語のパターンを表している。

上記の数値から言えることは，C1 の文書が全ての項目を含んでいることと，いくつかのコードは全ての文書に含まれているということである。

● C1 の結果

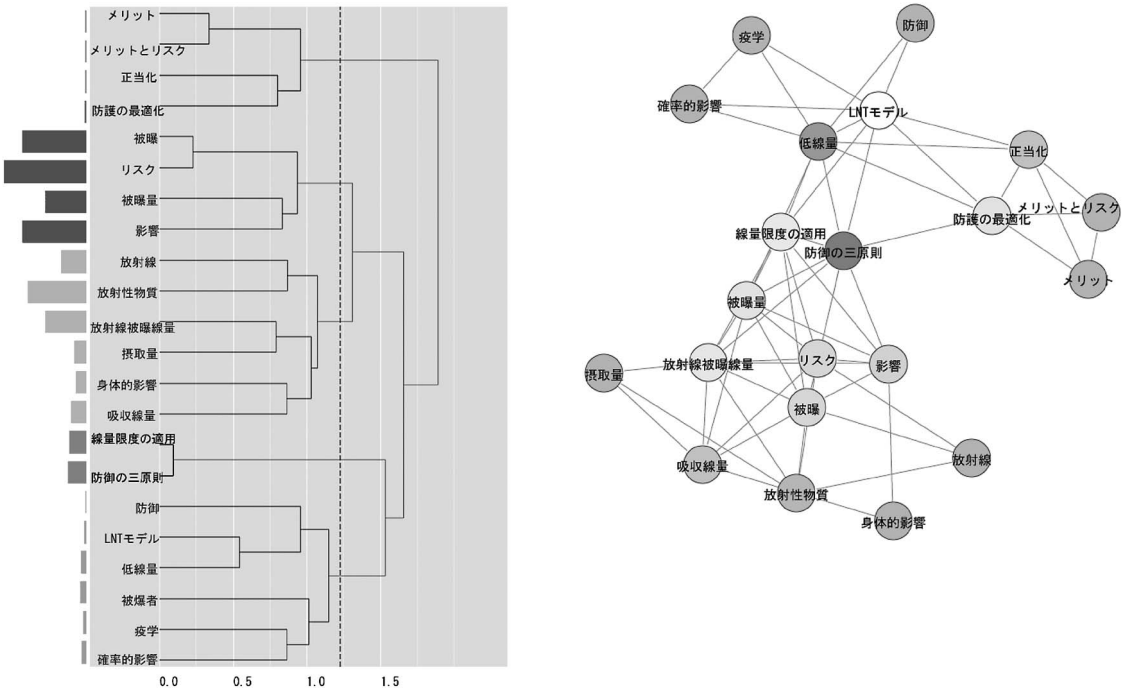

図-4　C1 のクラスター分析と共起ネットワーク

この文書には，全てのコードが含まれていることは，数値データでも明らかであったが。特に，被爆と放射線に関する記述が多いことが分かる。また，クラスタリングの結果が辞書のモデルのベン図と一致していることから，この図が文書の文脈を表していうことが分かる。この共起ネットワークのピンク（濃い灰色）および白になっているコードが媒体の中心性の高さを表している。それらが，低線量の放射線の安全性を考える上で非常に重要である LNT モデル，防衛の三原則をさしていることから，これらのコードが重要な役割を果たしていることが分かる。

● C2 の結果

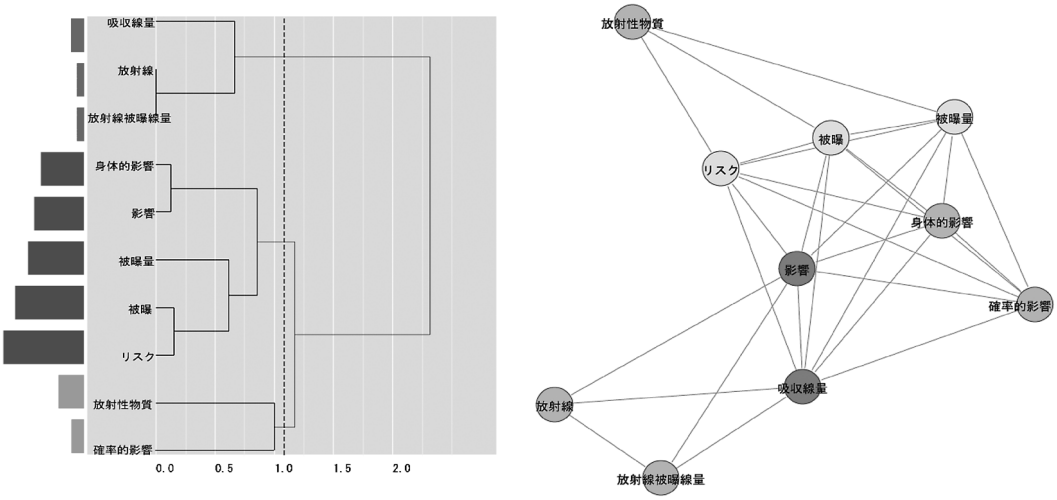

図-5　C2 のクラスター分析と共起ネットワーク

不安に対しての解説ということから，人体への影響について多く書かれていることが，この二つの図から分かる。

● C3 の結果

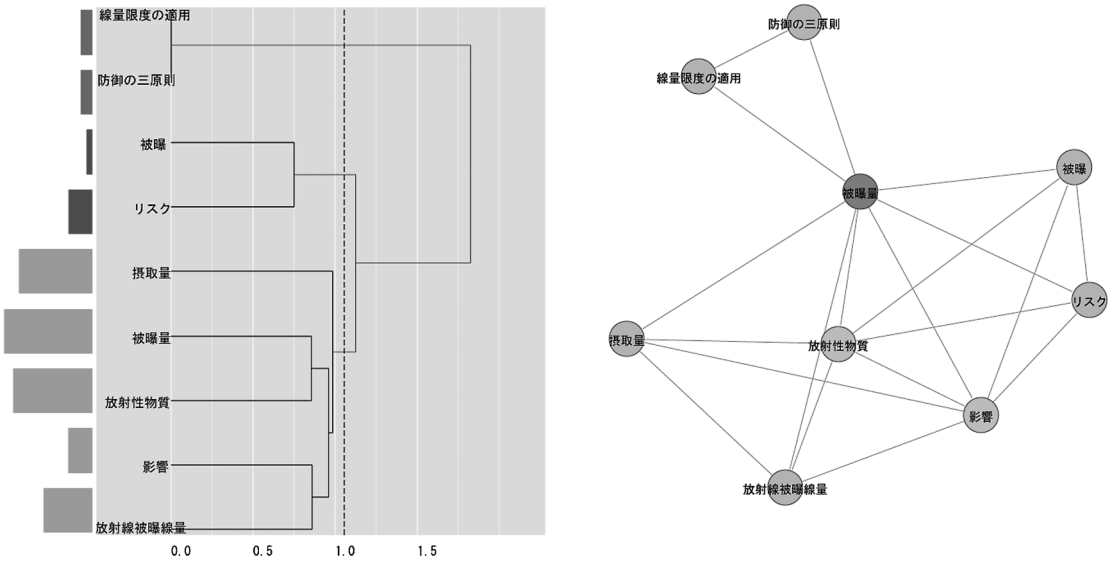

図-6　C3 のクラスター分析と共起ネットワーク

食品のガイドラインということから放射性物質と被曝量について重点的に解説していることが分かる。

● C4 の結果

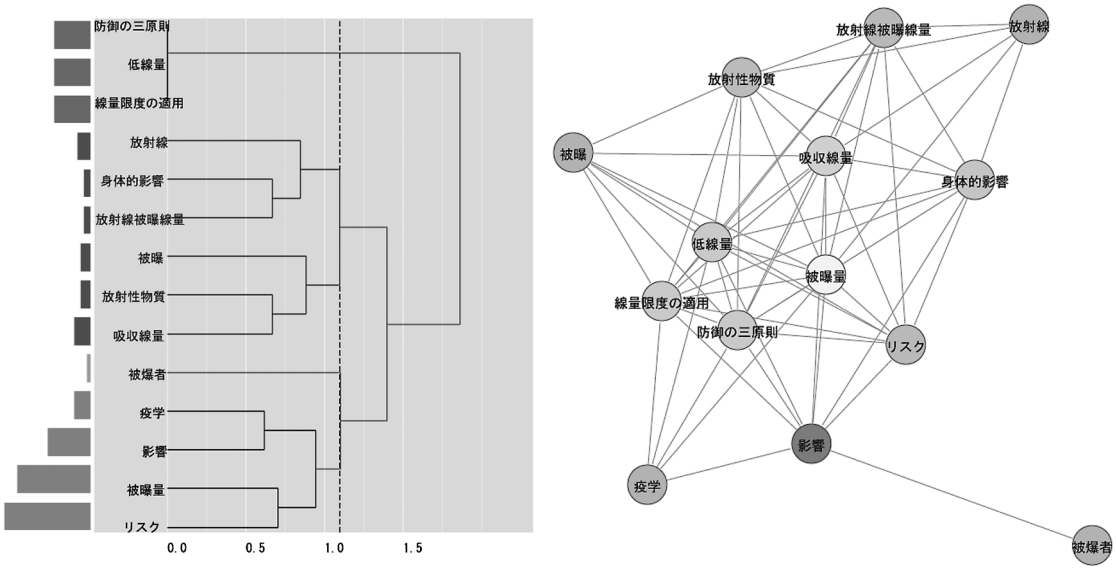

図-7　C4 のクラスター分析と共起ネットワーク

表題に不確実性をおいており，不確実な低線量の放射性の影響について，主要な論議の防御の三原則と被爆のリスクが多く記述されていることが分かる。

● C5 の結果

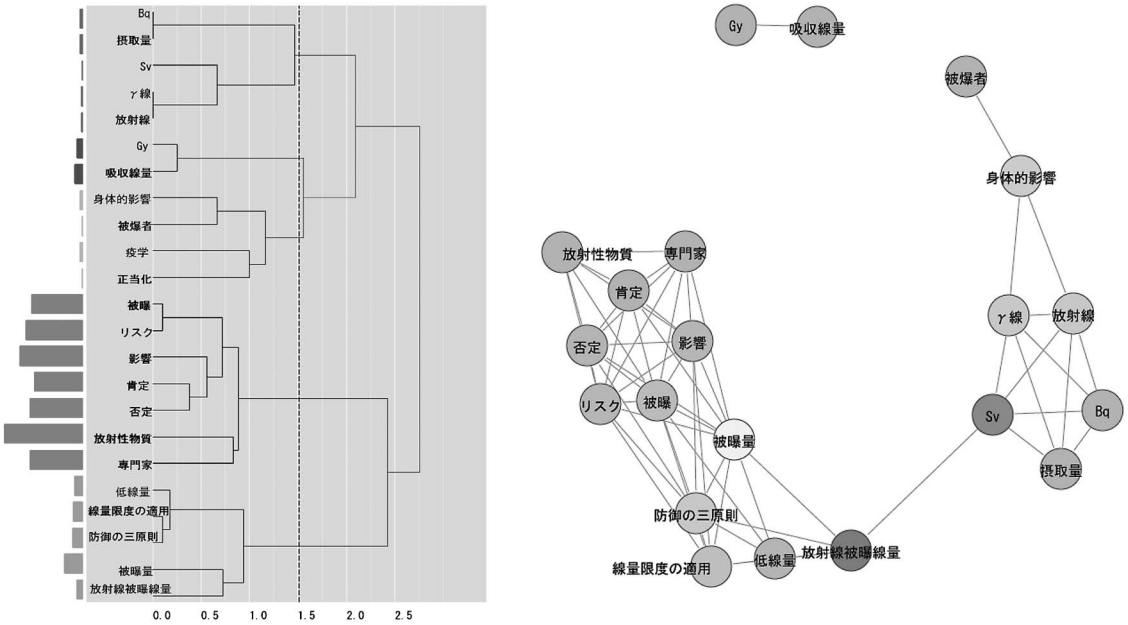

図-8　C5 のクラスター分析と共起ネットワーク

この文書は，コード数を増やさなければ，KH Coder がチャートを作成してくれなかったため，細かなコードも含めた図となっている。被爆のリスクについて多くの記述をしているのが特徴といえる。

● C6 の結果

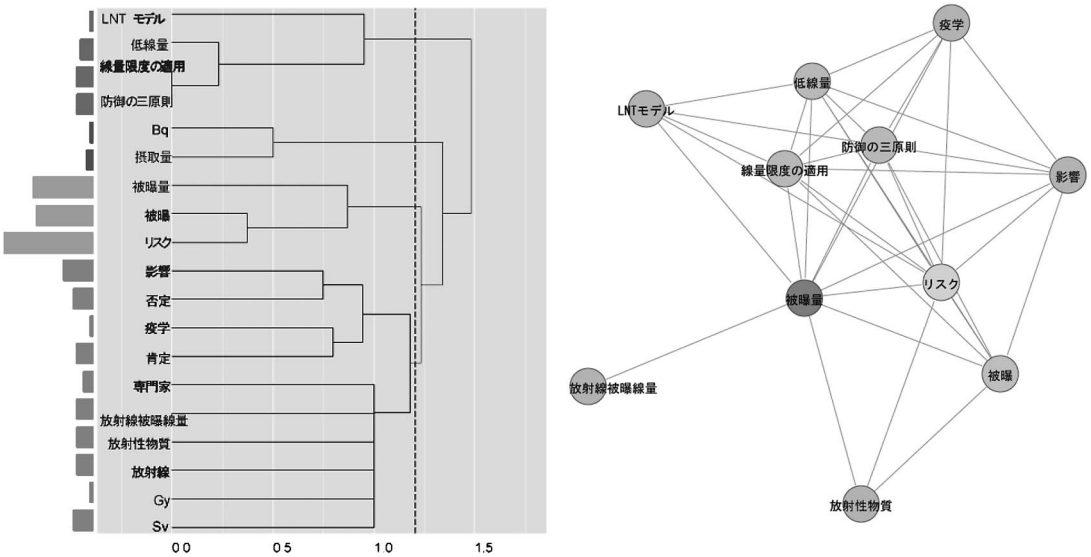

図-9　C6 のクラスター分析と共起ネットワーク

各国が安全放射性量の基準値としている ICPR 勧告についての解説で，やはり被爆とそのリスクについて多く記述していることが分かる。

● C7 の結果

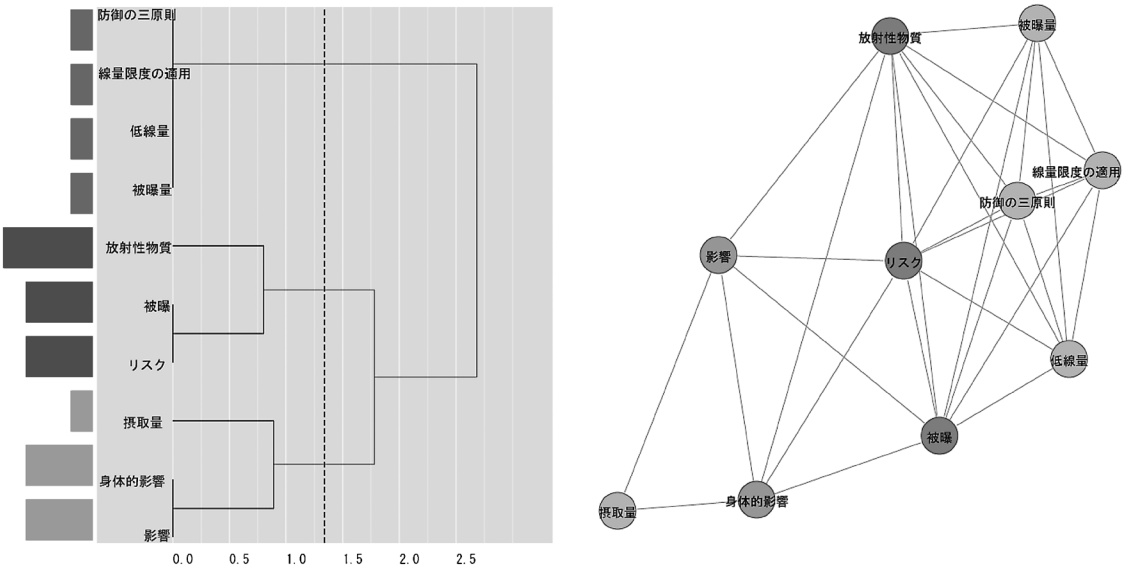

図-10　C7 のクラスター分析と共起ネットワーク

表題が示すように，被爆と身体的影響について多く記述されていうことが分かる。

● 7 つの文書のパターンの類似度に関する自己組織化マップ

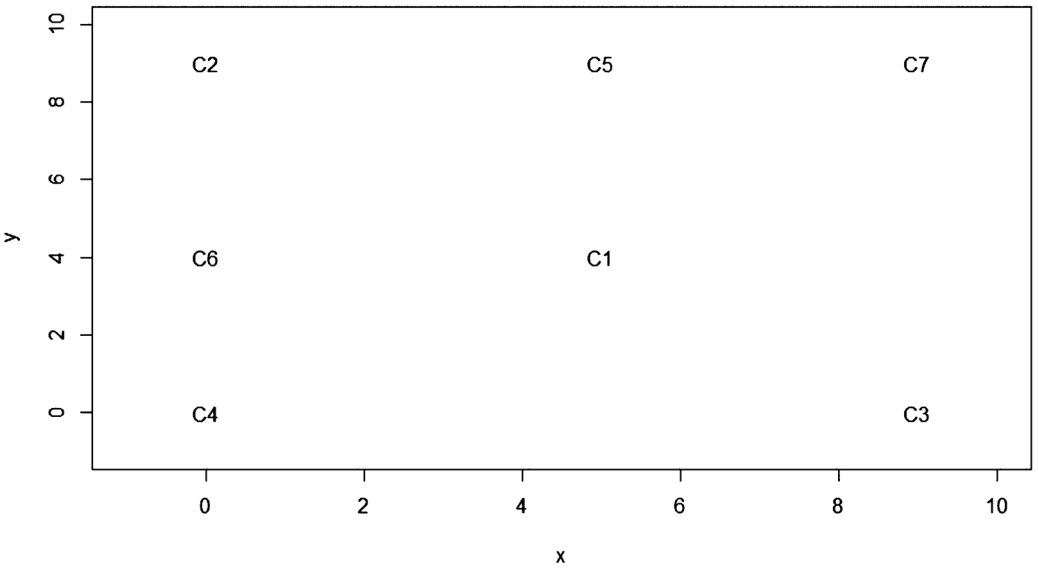

図-11　7 つの文書の自己組織化マップ

表 2 の 7 つの文書における主要コードの出現パーセンテージの数値のパターンを元に10000回の学習の結果がこの図である。図の x と y には特に意味は無い。測定者によって x と y の値域を設定されると，プログラムがそれぞれの計算対象の距離を元に，その xy 空間に対象の位置を設定する。つまり，意味があるのはそれぞれの対象における自分と自分以外の対象との距離である。

表2の7つの文書における主要コードの出現パーセンテージの数値のパターンを元に10000回の学習の結果がこの図である。

図のxとyには特に意味は無い。測定者によってxとyの値域を設定されると，プログラムがそれぞれの計算対象の距離を元に，そのxy空間に対象の位置を設定する。つまり，意味があるのはそれぞれの対象における自分と自分以外の対象との距離である。

このグラフでは，それぞれの対象の類似度が距離である。類似度は，表2で示した主要コードの出現数を元にするが，学習を重ねる度に類似している対象同士の距離を縮めていく。逆に類似度が少ないと距離が増加する。この測定では10000回の学習をおこなった。その結果，C1を中心に全ての対象が領域の限界に近い距離に離れている。つまり，今回の対象で類似度の高いものはないと言える。

しかし，マップの位置によりそれぞれの関係が分かる。最も網羅性の高い文書であるC1が全ての文書の中心にきている。つまり，情報の豊富さ故にすべての文章と似通った点をもっていることを表している。また，C2, C3とC4, C7の二組みが最も離れた文書といえる。C4, C7は，全く異なる種類の組織であるが，C2, C3は共に政府の文書である。だが，C2は国民の不安を低減することを目的とした文書であり，C3は食品の安全のためのガイドラインであるために異なるパターンをしめしていると思われる。その，C2からC3を線で結ぶとちょうど領域を2分する対角線となる。そして，その線で分けられたC4とC6，C5とC7が7つの文書の中では比較的近い存在である。C4とC6いずれも学術組織であるという共通点をもっている。また，C5は科学の専門家でありながら，報告書の冒頭に，客観的な事実ではなく，一人の人間の命の尊さを掲げる人間性を強調した文書となっている。また，C7の新聞記事の内容も自分の体験を元に人類に警告を与える人間性のある内容である。

5. 結　　論

今回のテキストマイニングでは，辞書作成に力点をおいて行った。その結果，放射線とその影響で尚且つ低線量に領域を限ったのに大変複雑な辞書となってしまった。

これは，放射線と言われるものが多種多様な実体を持つものの総称で極めて抽象的な言葉である事，また，放射線の人体への影響も実体は複雑で多様であることが原因と思われる。

また，専門家の解説も立場や場面によって使用されるキーワードが異なることが数値的にもチャートとしても結果として表すことができたと思える。

これらが，非専門家にとって放射線の解説が分り難いと感じる一因となっていると思われる。

参 考 文 献

［1］ 科学技術・学術政策研究所，「訪問面接方式による 科学技術に関する意識調査の結果ついて」，http://www.nistep.go.jp/wp/wp-content/uploads/201107_face-to-face_survey.pdf，（2011年7月）
［2］ 早川雄司，「東日本大震災後の国民の科学技術に関する意識の変化等について」，科学技術・学術政策研究所第2調査研究グループ，http://www.nistep.go.jp/wp/wp-content/uploads/review6-4.pdf，（2013年12月12日）
［3］ ICRP医療における放射線防護，Annals of the ICRP（ICRP年報）ICRP Publication 105医療における放射線防護，Elsevier社，（勧告，2007/10/1），http://www.icrp.org/docs/p105_japanese.pdf
［4］ EC European Committee on Radiation Risk，RR 2010 Recommendations of the European Committee on Radiation Risk，2010 Recommendations of the ECRR The Health Effects of Exposure to Low Doses of Ionizing Radiation，（勧告，2010/1/1, http://www.euradcom.org/2011/ecrr2010.pdf
［5］ 日本学術協力財団，「実りある不一致」のために　（特集　原発災害をめぐる科学者の社会的責任：科学と科学を

超えるもの）学術の動向：SCJフォーラム，（論文，2012/5/1）
http://www.scj.go.jp/ja/event/pdf/133-s-1-2.pdf
［6］ 厚生労働省，食品に含まれる放射性物質に関するガイドライン20130417-1，東日本大震災関連情報，（記事，2013/4/17），http://www.mhlw.go.jp/shinsai_jouhou/dl/20130417-1.pdf
［7］ The International Commission on Radiological Protection,（Annals of the ICRP ICRP PUBLICATION 115 Lung Cancer Risk from Radon and Progeny and Statement on Radon，ICRP Publication,（論文，2013/6/1）
http://www.sciencedirect.com/science/journal/01466453
［8］ 独立行政法人　産業技術総合研究所，今後の除染方針に関する議論の土台となる情報基盤を提供，放射性物質除染の効果と費用を評価，（報告書，2013/6/4）
http://www.aist.go.jp/aist_j/new_research/nr20130604/nr20130604.html
［9］ 環境放射能除染学会，Contamination of forests with radiocaesium − lessons from the Chernobyl accident，第2回環境放射能除染研究発表会 放射能除染のための国際シンポジウム，（論文，2013/6/7）
http://khjosen.org/2nd_Con/sympo_slide/Contamination%20of%20forests%20with%20radiocaesium-lessons%20from%20the%20Chernobyl%20accidentDr.George%20Shaw.pdf
［10］ 日本学術会議 社会学委員会 東日本大震災の被害構造と日本社会の再建の道を探る分科会，原発災害からの回復と復興のために必要な課題と取り組み態勢についての提言，日本学術会議社会学委員会　東日本大震災の被害構造と日本社会の再建の道を探る分科会，（論文，2013/6/27）
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t174-1.pdf
［11］ 小林傳司，2007/6/27，トランス・サイエンスの時代　科学技術と社会を繋ぐ，NTT 出版株式会社
［12］ 那須川哲哉，諸橋正幸，長野　徹，1999/7/16，膨大な文書データの自動分析による知識発見，情報処理学会研究報告．DD，一般社団法人情報処理学，99(57)，65−72，http://ci.nii.ac.jp/naid/110004029314
［13］ 橋本泰一，村上浩司，乾　孝司，内海和夫，石川正道，2007/5/31，共起語に基づいた階層型文書クラスタリング手法，社団法人情報処理学会研究報告，社団法人情報処理学会，54，13−20，
http://ci.nii.ac.jp/naid/110006292333
［14］ 橋本泰一，村上浩司，乾　孝司，内海和夫，石川正道，2008/3/1，文書クラスタリングによるトピック抽出および課題発見，社会技術研究論文集，社会技術研究会，Vol. 5，216−226，https://www.jstage.jst.go.jp/article/sociotechnica/5/0/5_0_216/_article/-char/ja/
［15］ 服部　峻，大島裕明，小山　聡，田中克己，2007/7/2，語の同位関連と性質の継承関連を用いた概念階層のWeb からの抽出（夏のデータベースワークショップ2007（データ工学，一般）），情報処理学会研究報告．データベース・システム研究報告会，一般社団法人情報処理学会，2007(65)，127〜132，　http://ci.nii.ac.jp/naid/110006381408
［16］ 三末和夫，渡部　勇，1999/7/16，テキストマイニングのための連想関係の可視化技術，情報処理学会研究報告．DD，一般社団法人情報処理学会，99(57)，65−72，http://ci.nii.ac.jp/naid/110004029314
［17］ 盧世　淼，峯　恒憲，雨宮真人，2002/3/4，国際会議の論文募集ファイルからのトピックスの抽出とクラスタリング，情報処理学会研究報告．自然言語処理研究報告，一般社団法人情報処理学会，2002(20)，9〜16，http://ci.nii.ac.jp/naid/110002935395
［18］ 山本真照，2011/4/1，テキストマイニング手法の洗練に向けた知識活用方法に関する研究，経済科学論究，埼玉大学経済学会，第8号，73−85，
http://sucra.saitama-u.ac.jp/modules/xoonips/detail.php?id＝KY−AA11950211−08−08
［19］ 秋光淳生：データからの知識発見，288ページ，放送大学教育振興会，2012/03
［20］ 押切孝雄：グーグル・マーケティング!，2008/6/19, pp. 25，技術評論社，2008/6/19
［21］ 石田基広：R によるテキストマイニング入門，2008/12/16, pp. 192，森北出版，2008/12/16
［22］ 長谷川豊：トランス・サイエンス領域でのインターネットサービスについて，Internet Services for the Area of Trans-Sciences，国士舘大学紀要情報科学(35)，24−35，(2014)