（２）事故情報データの品質向上に向けて（村田 磨理子）

ア　目的

事故情報データバンクに登録されている事故情報に関して、収録されている項目や分類の状況を確認することにより、データの品質を検証する。検証結果からデータの有用性を高める方策を検討する。

データの品質を向上させることにより、

- これまで、（項目指定）検索・集計に利用されていなかった項目について、適切に分類して検索・集計を可能とすることにより、利用者ニーズの適合性が向上し、利用の拡大が期待される。
- 項目の定義や分類基準を明確にすることによって、事故情報集計の明確性や比較可能性を向上させ、さらに統計分析を促進することが期待される。
- 一定の標準的な項目や分類基準を示すことにより、より多くの消費者からの情報提供が容易になることや、情報入力担当者の負担軽減が期待され、データバンクの効率性の向上に寄与することができる。

イ　分析

データの収録内容について、度数分布表、クロス集計表やグラフを用いて、項目や分類に関するニーズ適合性や整合性・比較可能性を検証した。分析対象のデータは次のとおり。

- 83項目（一般向けウェブサイトでは非表示の項目を含む）
- 198,135レコード（登録年月日が2009年10月1日から2016年12月31日まで）

本報告は、既存の完成されたデータファイルの内容を利用した分析の範囲にとどまるが、分析結果からは、データのコード化及び編集に直接関係するもののほかに、設計やデータの収集などに関する示唆も得られた。

（ア）項目の収録状況

事故情報データバンクは、各機関で持っているデータベースの項目に基づいて事故情報を入力することを基本としているため、参画機関ごとに入力項目に差異が生じている。このため、参画機関をとおして共通して入力することとされている項目が限定されている。そこで、研究者、行政機関等利用者、一般利用者などのニーズを把握して、入力を推奨する項目を明確にすべきと考える。

本専門調査会において、自由記述の有用性を高める枠組みなどが提案されており、項目の選択にも役立つと考えられる。

分析に使用したデータにおいては、項目によっては、空白が非常に多く、全レコードの95%以上が空白である項目は40項目、そのうち13項目はすべて空白であった。空白の原因は、無記入、該当なしなどの理由が考えられるが、利用にあたっての説明が必要だと考える。

（イ）収録の形式

発生日時を例にとり、収録された内容の度数分布表を作成して確認すると、次のように様々な形式になっていることが分かる。

- 年のみ
- 年月
- 年月日
- 年月日時
- 年月日時分
- 年月時など
- 複数の時点を併記
- 季節
- 期間
- 頻度
- 発生条件（例：「夏 気温30度以上」）
- 現在または基準となる時点からの経過・遡及時間（例：「購入から１年」）

事故情報のトレンドをみるとき、データの収録項目として、発生時点、受付時点、データバンクへの登録時点の３つを基準として時系列分析を行うことが想定される。しかし、発生時点は空白が約３割あることに加えて上記のように形式がまちまちであるため、使い勝手がよくない。

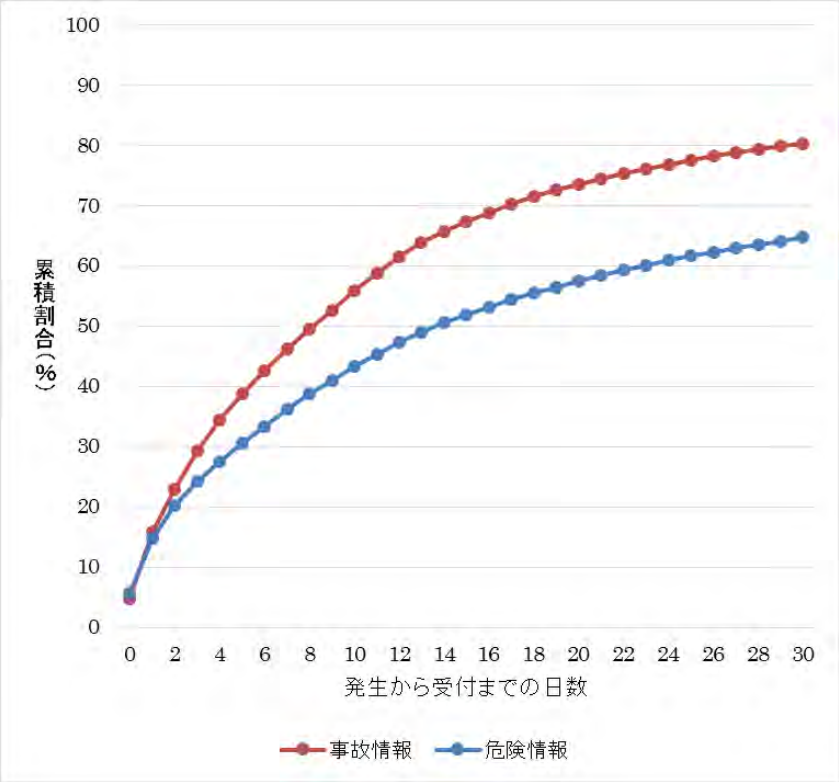
改善策として、まず、年月日と時刻は定型フォーマットとし、その他の自由記述文と分けて収録されることが望ましい。また、空白を減らすためには、最初の情報提供を受ける際に、少なくとも「年」を必須とするなどの対応が併せて必要だろう。

なお、現状のデータに対して形式統一処理をしたとすると、半数は「年月日」が利用でき、約６割は「年月」の情報が利用できると思われる。

ここでは、発生日時と受付日時の両方から年月日を取り出せた一部のデータ（82,698件）を使って、事故等の発生から受付までの日数を集計した結果を示す。（ただし、日数が利用できないものが約10万件あるため、本当の経過日数の分布を正確に推測することは困難である点に留意する必要がある。）

事故情報は発生から30日までに８割が受付され、危険情報は30日で６割程度が受付される。中央値で比較すると、事故情報は９日、危険情報は14日である。

【図表２－６】発生から受付までの日数別件数の累積割合



この結果は、受付時点からみると、例えば、今受付の件数が増えているということが、今発生が増えていることと必ずしも同じではないことを示唆している。特に危険情報は、４割近くは１か月以上前に発生した案件であった。なお、発生から時間が経っている案件は、類似案件の報道発表に触発されて通報するといった行動が考えられる。受付件数の増減は、報道発表との関連を併せてみることが重要と思われる。

次に、事故情報に限定して、情報提供元（参画機関）ごとにみると、

日数の分布に違いがみられる。違いは、事故の被害者が情報提供しているのか、それとも、製品・サービスに係る事業者からの情報提供なのかといった、通報者の属性、通報の手段・経路に起因すると推察するが、現状のデータの範囲では明らかにできなかった。違いの要因となる項目をデータに追加することで、分析が深まると期待される。少なくとも、データの利用者に対して、参画機関ごとの特徴を説明することが望ましいのではないか。

（ウ）クロス集計

被害者の年代、性別といった属性と、商品や事故内容のクロス集計によって、事故の特徴を分析したいが、現状のデータでは、項目ごとの空白の多さや収録形式の課題があり、クロス集計が容易ではない。

一例として、被害者性別の空白が比較的少ない情報提供元に限定して、被害者性別と商品の分類とのクロス集計を試みた。

【図表２－７】被害者性別と商品分類とのクロス集計

	男性	女性	不明、空白
食料品	6,975	12,535	7,793
家電製品	1,133	1,628	8,616
住居品	3,082	5,789	8,287
文具・娯楽用品	1,339	1,517	1,718
光熱水品	187	260	1,125
被服品	875	2,247	595
保健衛生品	4,831	26,468	2,136
車両・乗り物	1,618	1,355	8,051
建物・設備	2,194	4,362	2,860
保健・福祉サービス	3,881	17,445	667
他の商品・サービス	3,959	6,659	5,117
総数	22,993	60,750	37,620

（注）商品など分類は複数に該当するため、内訳の和は総数に一致しない。

総数では、男性22,993人、女性60,750人、不明・空白37,620人であり、女性が男性の2.6倍となっている。商品など分類ごとにみると、男女比の偏りが小さいものから大きいものまでさまざまであり、保健衛生品及び保健・福祉サービスにおける女性の多さが際立っている。

一方で、この結果において、女性が男性の2.6倍となっていることが、

全体で発生している事故の比率と同じとみることは危険である。実際に発生している事故等のうち、情報提供される割合が偏っている可能性が高く、バイアスがあると考えられる。

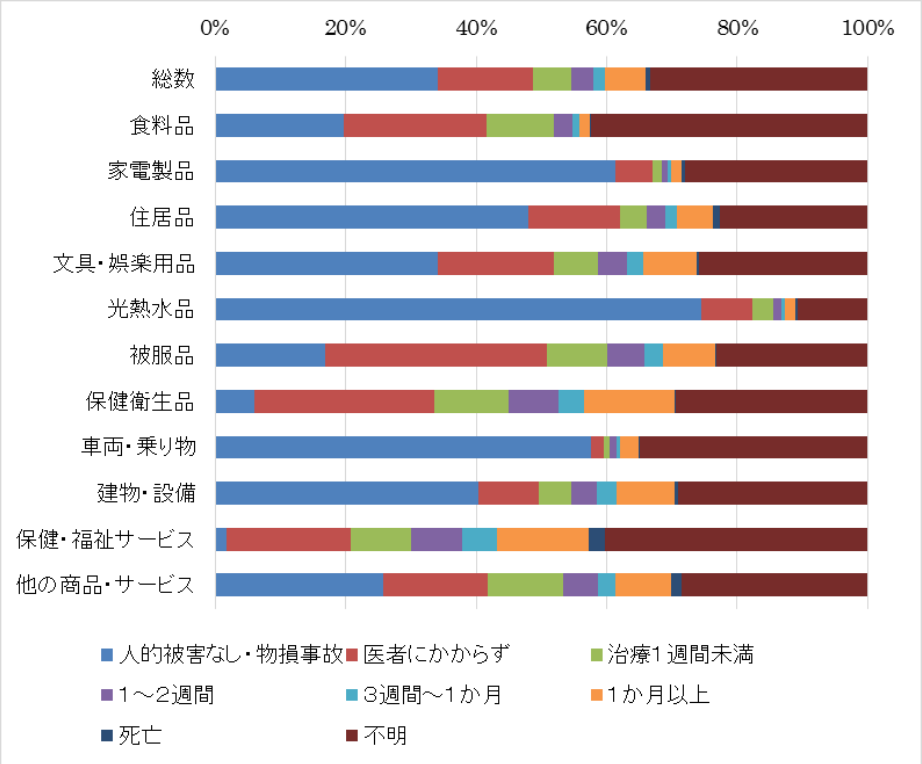
（エ）項目の区分

事故情報には傷病の程度の収録が比較的多いが、危険情報ではほぼ空白となっている。危険情報の定義は「けが人が発生していないものの、発生するおそれがあった事案等に関する情報」であるため、傷病の程度が空白であると考えられるが、「人的被害なし・物損事故」といった区分を追加して、事故情報と一貫した順序尺度をもつ区分とすることによって、傷病の程度の分析の幅が広がると思われる。

ここでは、危険情報の傷病の程度をすべて「人的被害なし・物損事故」とみなして、事故情報と併せて、商品など分類別に集計した結果を示す。

保健・福祉サービス、保健衛生品、被服品、食料品は人的被害が出ている割合が高く、保健・福祉サービス、保健衛生品は１か月以上・死亡の割合も比較的高いこと、光熱水品、家電製品、車両・乗り物は人的被害なし・物損事故の割合が高いことが示される。

【図表２－８】商品など分類別傷病の程度の構成割合



ただし、この結果についても、前述のとおり、全体で発生している事故の傷病の程度と同じとみることは危険であると考える。情報提供される割合が商品など分類によって偏っている可能性がある。

（オ）自由記述の扱いについて

情報提供元によっては、商品や事故内容の分類ルールを詳細に設定していることを確認した。しかし、複数の情報提供元の間で、分類ルールが統一されているかは不明であり、データをそのまま一括して使うと、分類の揺らぎが表面化していると推察する。

そこで、事故情報データバンクシステムにおいて、統一的なアフターコード化[19]を実施することを提案する。（公的統計の作成では、例えば、独立行政法人統計センターにおいて、経済センサスの産業分類や社会生活基本調査の生活行動分類などについて、自由記述文から自動的にコーディングする方法が研究され、実用化されている。）本専門調査会においても、言語解析技術や機械学習を活用した分析を行っており、同技術を活用することで、統一的なアフターコード化を確立できると思われる。

分類の揺らぎのほかに、適切な分類区分を網羅しているかという点からみると、例えば、事故内容は、「その他」が４割を占める。「その他」の割合は、情報提供元によって、大きく異なる。そこで、情報提供元による分類の違いを考慮に入れて、統一的なアフターコード化により、「その他」の細分化を検討する余地があると思われる。

また、現状では空白が多い事故原因、措置状況に相当する内容が、別の項目の自由記述文に含まれる場合や、関連する２つの項目が必ずしも整合していない場合などが散見されており、統一的なアフターコード化の対象とすることにより、空白を減らし整合性のある有用性の高いデータとすることが期待される。

（カ）データ収録内容に関するその他の課題

次に挙げる点を解消することで、データの有用性を高めることができると思われる。

- 「商品など分類」のように、複数の区分に該当するような場合のデータ構造を工夫する。（集計用のデータでは、区分ごとに「該当す

[19] 定型化された言葉や記号（コード）ではない状態から、事後的に（アフター）記号を割り当てる（コード化、コーディング）ことをアフターコード化という。対して、選択肢のように事前に記号（コード）を用意したものはプレコードという。

る、しない」のフラグを入れる方法がよく使われている。）

- 「商品など分類」の大分類、中分類のように異なるレベルで分類するものは、大分類と中分類をそれぞれ別の項目として収録する。
- 「被害者の情報　年代等－年代」の年齢階級は、利用者ニーズと適合しているか確認する。
- 「被害者の情報　年代等－属性」の使い方を明確にする。（現状の収録内容から推察すると、「要介護者」といった情報を補足するものであることを明確にするなど）
- 「原因究明等－原因調査状況」と「原因究明等－事故原因」の項目間の関連を明確にする。（例えば、「原因究明等－事故原因」に記載があれば、「原因究明等－原因調査状況」は空白にしないなど）
- 「発生場所」、「発生場所－施設用途」、「発生場所－施設名」、「発生場所－発生場所」は、空白がかなり多く、４つの項目の使い分けが難しいと感じる。特に、「発生場所－施設名」は、「発生場所－施設用途」との混同が多い。
- 「発生場所」に番地までの住所が必要かどうか。

ウ　今後の課題

分析結果から得られた、データの品質向上のための課題を下記にまとめる。

- 自由記述文とコード化された項目を明確に分ける。
- 自由記述文からの分類、コード化は情報提供元による定義の違いを考慮する。可能ならば、事故情報データバンクシステム独自のアフターコード化を行う。
- 項目間の関連性を明確に定義する。
- 利用者のニーズに応じた情報提供として、事故原因、措置状況、被害者の性別など、現状では実質的に利用できない項目について、ニーズに応じてさらなる提供を検討する。
- データの説明文書を充実する。（事故情報登録作業向けと、データ利用者向け）

（３）事故情報の活用等のあり方について　言語解析技術（相澤 彰子）

ア　目的

事故情報データバンクに登録されている事故情報に関して、言語解析技術を活用した分析を行うことにより、分類項目の網羅性や整合性を調査する。また、自由記述項目に対する言語解析の適用可能性を調べ、効率的なデータ収集・管理・分析に向けた課題の整理を行う。

イ　分析する事故情報の分野

まず、事故情報データバンクに登録されたすべてのレコード198,135件について、俯瞰的な調査を実施する（以下、「分類項目ごとのデータ登録状況に関する俯瞰的な調査」と呼ぶ)。次に、レコード件数や自由記述項目に登録されたテキストの量を勘案し、《情報提供元》が《国民生活センター》で、《事故種別》が《事故情報》である85,778件をサンプルとして抽出し、言語解析を含む詳細な分析を行う（以下、「言語解析の適用可能性に関する調査」と呼ぶ)。

ウ　分析に用いられた技術の概要

（ア）分類項目ごとのデータ登録状況に関する俯瞰的な調査

全事故情報 198,135件を Unicode（UTF-８）に変換して、83個の分類項目について、「辞書サイズ」、「網羅率」、「平均バイト数」を調べた。ここで「辞書サイズ」とは、項目ごとの登録文字列の異なり数とする。たとえば分類項目《種別》に対しては、《事故情報》と《危険情報》の２つの値が登録されていることから、辞書サイズは２となる。分類項目《事故内容詳細》のように自由記述項目である場合には、登録文字列は原則として互いに異なるため、辞書サイズはレコード数に近くなる。また、「網羅率」とは、値が登録されていない、いわゆる欠損値を除くデータの割合と定義する。「平均バイト数」は、項目ごとに登録されているテキストの長さの平均で、欠損値を除いて計算する。調査はUTF-８で符号化されたテキストを対象に行ったため、日本語については１文字３バイトで計算される。

調査の結果、分類項目ごとのデータ登録状況は、情報提供元によって大きく異なることが判明したため、情報提供元URLが入力されていて登録件数が1,000件以上ある５つの情報提供元について、それぞれの数値を求め、図表２－９のとおり比較可能な形に整理をした。

【図表２－９】分類項目ごとのデータ登録状況比較

		全体			参画機関①			参画機関②		
番号	フィールド名	辞書サイズ	網羅率	平均バイト数	辞書サイズ	網羅率	平均バイト数	辞書サイズ	網羅率	平均バイト数
0	No	9,939	1.00	3.8	9,939	1.00	3.8	5,824	1.00	4.0
1	事故情報ID	198,135	1.00	9.0	121,363	1.00	8.9	8,308	1.00	9.2
2	事故概要	42,985	1.00	22.5	30,651	1.00	17.3	5,442	1.00	52.5
3	種別	2	1.00	12.0	2	1.00	12.0	1	1.00	12.0
4	発生日時	19,250	0.65	16.6	14,097	0.49	16.4	2,994	1.00	16.8
5	発生場所	219	0.46	9.4	48	0.49	9.3	47	0.87	9.3
6	商品など分類	2,177	0.98	42.1	2,170	1.00	49.5	1	1.00	27.0
7	商品など名称	55,808	1.00	29.5	43,355	1.00	24.2	5,459	1.00	70.1
8	事故内容	20	0.92	13.0	17	1.00	13.3	1	1.00	12.0
9	事故内容詳細	167,882	0.94	610.2	121,316	1.00	865.6	0	0.00	0.0
10	傷病内容	28	0.52	18.2	20	0.71	19.6	1	1.00	9.0
11	傷病の程度	8	0.45	15.3	7	0.71	15.4	0	0.00	0.0
12	措置状況	3	0.06	9.1	0	0.00	0.0	0	0.00	0.0
13	情報提供元	13	1.00	27.4	1	1.00	24.0	1	1.00	15.0
14	情報提供元URL	7	0.88	37.8	1	1.00	39.0	1	1.00	38.0
15	登録年月日	2,432	1.00	16.7	2,387	1.00	16.7	48	1.00	16.8
16	備考	125,713	0.74	424.0	116,475	0.99	506.9	0	0.00	0.0
17	通知方法	6	0.65	6.1	3	0.97	6.0	0	0.00	0.0
18	被害者の情報-人数	277	0.10	4.3	0	0.00	0.0	247	1.00	4.5
19	被害者の情報-氏名	5	0.00	9.6	0	0.00	0.0	0	0.00	0.0
20	被害者の情報-所属-職業	20	0.62	11.9	10	1.00	11.8	0	0.00	0.0
21	被害者の情報-所属-住所	197	0.42	9.4	49	0.69	9.3	0	0.00	0.0
22	被害者の情報-年代等-年代	51	0.41	8.2	11	0.64	8.2	0	0.00	0.0
23	被害者の情報-年代等-属性	6	0.00	9.3	0	0.00	0.0	0	0.00	0.0
24	被害者の情報-性別	3	0.45	6.0	3	0.70	6.0	0	0.00	0.0
25	発生場所-施設用途	41	0.53	12.5	9	0.67	11.8	24	1.00	20.2
26	発生場所-施設名	62	0.00	14.0	0	0.00	0.0	0	0.00	0.0
27	発生場所-発生場所	141	0.08	9.4	10	0.12	7.3	0	0.00	0.0
28	型式・ロット・生産国	68,282	0.48	57.2	32,571	0.36	32.1	0	0.00	0.0
29	傷病内容詳細	340	0.01	24.4	0	0.00	0.0	0	0.00	0.0
30	傷病の程度詳細	296	0.09	18.7	0	0.00	0.0	0	0.00	0.0
31	事業者-事業者名1	47,913	0.69	21.7	39,887	0.85	18.0	0	0.00	0.0
32	事業者-事業区分1	1,304	0.12	18.4	1,294	0.02	44.2	0	0.00	0.0
33	事業担当部署-所在地1	5	0.00	47.0	0	0.00	0.0	0	0.00	0.0
34	事業担当部署-電話番号1	5	0.00	11.6	0	0.00	0.0	0	0.00	0.0
35	事業担当部署-担当部署1	3	0.00	20.0	0	0.00	0.0	0	0.00	0.0
36	事業担当部署-担当者役職1	2	0.00	7.5	0	0.00	0.0	0	0.00	0.0
37	事業担当部署-氏名1	3	0.00	8.0	0	0.00	0.0	0	0.00	0.0
38	製造・輸入時期1	2	0.00	15.5	0	0.00	0.0	0	0.00	0.0
39	購入・販売時期1	5,087	0.45	13.0	5,084	0.73	13.0	0	0.00	0.0
40	設置期間1	1	0.00	39.0	0	0.00	0.0	0	0.00	0.0
41	使用期間1	425	0.23	10.8	37	0.22	13.5	0	0.00	0.0
42	事業者-事業者名2	9,468	0.14	17.1	9,052	0.22	16.9	0	0.00	0.0
43	事業者-事業区分2	119	0.00	19.2	111	0.00	37.9	0	0.00	0.0
44	事業担当部署-所在地2	3	0.00	45.7	0	0.00	0.0	0	0.00	0.0
45	事業担当部署-電話番号2	1	0.00	12.0	0	0.00	0.0	0	0.00	0.0
46	事業担当部署-担当部署2	0	0.00	12.0	0	0.00	0.0	0	0.00	0.0
47	事業担当部署-担当者役職2	1	0.00	15.0	0	0.00	0.0	0	0.00	0.0
48	事業担当部署-氏名2	1	0.00	15.0	0	0.00	0.0	0	0.00	0.0
49	製造・輸入時期2	0	0.00	15.0	0	0.00	0.0	0	0.00	0.0
50	購入・販売時期2	0	0.00	15.0	0	0.00	0.0	0	0.00	0.0
51	設置期間2	0	0.00	15.0	0	0.00	0.0	0	0.00	0.0
52	使用期間2	1	0.00	11.0	0	0.00	0.0	0	0.00	0.0
53	事業者-事業者名3	16	0.00	20.8	0	0.00	0.0	0	0.00	0.0
54	事業者-事業区分3	6	0.00	16.1	0	0.00	0.0	0	0.00	0.0
55	事業担当部署-所在地3	0	0.00	16.1	0	0.00	0.0	0	0.00	0.0
56	事業担当部署-電話番号3	0	0.00	16.1	0	0.00	0.0	0	0.00	0.0
57	事業担当部署-担当部署3	0	0.00	16.1	0	0.00	0.0	0	0.00	0.0
58	事業担当部署-担当者役職3	0	0.00	16.1	0	0.00	0.0	0	0.00	0.0
59	事業担当部署-氏名3	0	0.00	16.1	0	0.00	0.0	0	0.00	0.0
60	製造・輸入時期3	0	0.00	16.1	0	0.00	0.0	0	0.00	0.0
61	購入・販売時期3	0	0.00	16.1	0	0.00	0.0	0	0.00	0.0
62	設置期間3	0	0.00	16.1	0	0.00	0.0	0	0.00	0.0
63	使用期間3	0	0.00	16.1	0	0.00	0.0	0	0.00	0.0
64	措置状況詳細	6,961	0.15	231.2	0	0.00	0.0	0	0.00	0.0
65	原因究明等-原因調査状況	5	0.83	2.2	1	1.00	1.0	1	1.00	1.0
66	原因究明等-事故原因	31	0.13	27.5	0	0.00	0.0	0	0.00	0.0
67	原因究明等-事故原因詳細	24,992	0.21	348.0	0	0.00	0.0	0	0.00	0.0
68	原因調査機関-機関名	6	0.00	32.2	0	0.00	0.0	0	0.00	0.0
69	原因調査機関-連絡先	11	0.00	25.3	0	0.00	0.0	0	0.00	0.0
70	通報者情報-氏名	12	0.00	20.5	0	0.00	0.0	0	0.00	0.0
71	通報者情報-属性	753	0.43	17.0	1	0.55	15.0	0	0.00	0.0
72	通報者情報-職業	19	0.61	13.7	12	1.00	13.7	0	0.00	0.0
73	通報者情報-住所	78	0.67	9.3	50	1.00	9.3	0	0.00	0.0
74	通報者情報-連絡先-TEL	5	0.00	16.6	0	0.00	0.0	0	0.00	0.0
75	通報者情報-連絡先-FAX	3	0.00	19.3	0	0.00	0.0	0	0.00	0.0
76	通報者情報-連絡先-E-MAIL	2	0.00	31.5	0	0.00	0.0	0	0.00	0.0
77	最終更新年月日	2,487	1.00	16.7	2,318	1.00	16.7	60	1.00	16.8
78	受付年月日	21,972	0.90	19.9	18,783	1.00	21.4	0	0.00	0.0
79	情報元システムID	198,101	1.00	13.6	121,363	1.00	16.0	8,308	1.00	11.0
80	情報提供機関名、担当課名-所属	1,665	0.88	30.6	1,505	1.00	38.1	120	1.00	9.8
81	情報提供機関名、担当課名-氏名	25	0.00	15.0	0	0.00	0.0	0	0.00	0.0
82	情報提供機関名、担当課名-TEL	23	0.00	9.6	0	0.00	0.0	0	0.00	0.0

		参画機関③			参画機関④			参画機関⑤		
番号	フィールド名	辞書サイズ	網羅率	平均バイト数	辞書サイズ	網羅率	平均バイト数	辞書サイズ	網羅率	平均バイト数
0	No	5,955	1.00	4.0	7,009	1.00	3.8	4,911	1.00	3.8
1	事故情報ID	20,060	1.00	9.5	17,653	1.00	8.8	7,901	1.00	9.0
2	事故概要	1,992	1.00	42.8	4,761	1.00	27.5	1,274	1.00	25.5
3	種別	2	1.00	12.0	2	0.99	12.0	1	1.00	12.0
4	発生日時	4,366	0.98	15.6	3,165	0.86	16.8	2,906	1.00	16.7
5	発生場所	0	0.00	0.0	65	0.97	9.4	47	1.00	9.3
6	商品など分類	25	0.84	33.9	56	1.00	31.8	32	1.00	30.5
7	商品など名称	2,016	1.00	48.7	4,783	1.00	28.2	1,280	1.00	25.7
8	事故内容	2	1.00	10.9	20	1.00	14.3	12	0.90	12.0
9	事故内容詳細	11,800	1.00	107.4	7,780	0.79	96.5	4,858	1.00	113.8
10	傷病内容	0	0.00	0.0	23	0.37	11.3	14	0.10	11.0
11	傷病の程度	2	0.01	10.4	7	0.08	13.7	2	0.14	12.6
12	措置状況	0	0.00	0.0	2	0.31	9.1	3	0.72	9.1
13	情報提供元	1	1.00	40.0	1	1.00	12.0	1	1.00	48.0
14	情報提供元URL	1	1.00	27.0	1	1.00	38.0	1	1.00	47.0
15	登録年月日	28	1.00	16.9	583	1.00	16.7	710	1.00	16.7
16	備考	0	0.00	0.0	1,109	1.00	49.5	7,901	1.00	15.1
17	通知方法	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
18	被害者の情報-人数	0	0.00	0.0	154	0.41	4.3	10	0.29	4.0
19	被害者の情報-氏名	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
20	被害者の情報-所属-職業	0	0.00	0.0	16	0.00	15.3	0	0.00	0.0
21	被害者の情報-所属-住所	0	0.00	0.0	30	0.01	9.5	0	0.00	0.0
22	被害者の情報-年代等-年代	0	0.00	0.0	41	0.13	8.7	11	0.02	8.8
23	被害者の情報-年代等-属性	0	0.00	0.0	1	0.00	24.0	0	0.00	0.0
24	被害者の情報-性別	0	0.00	0.0	3	0.13	6.0	3	0.02	6.0
25	発生場所-施設用途	0	0.00	0.0	10	0.79	12.2	0	0.00	0.0
26	発生場所-施設名	0	0.00	0.0	50	0.01	13.6	0	0.00	0.0
27	発生場所-発生場所	0	0.00	0.0	24	0.01	10.3	0	0.00	0.0
28	型式・ロット・生産国	3,995	0.65	37.2	9,201	0.66	27.2	4,342	0.68	15.0
29	傷病内容詳細	0	0.00	0.0	171	0.01	18.8	0	0.00	0.0
30	傷病の程度詳細	1	0.20	6.0	186	0.11	12.0	66	0.28	10.5
31	事業者-事業者名1	1,357	0.65	32.0	6,414	0.87	30.5	1,464	0.67	42.2
32	事業者-事業区分1	0	0.00	0.0	8	0.86	15.3	3	0.66	15.0
33	事業担当部署-所在地1	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
34	事業担当部署-電話番号1	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
35	事業担当部署-担当部署1	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
36	事業担当部署-担当者役職1	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
37	事業担当部署-氏名1	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
38	製造・輸入時期1	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
39	購入・販売時期1	0	0.00	0.0	2	0.00	17.0	0	0.00	0.0
40	設置期間1	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
41	使用期間1	385	1.00	7.4	0	0.00	0.0	0	0.00	0.0
42	事業者-事業者名2	0	0.00	0.0	468	0.04	27.9	2	0.00	33.0
43	事業者-事業区分2	0	0.00	0.0	8	0.04	15.5	1	0.00	15.0
44	事業担当部署-所在地2	0	0.00	0.0	1	0.00	27.0	0	0.00	0.0
45	事業担当部署-電話番号2	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
46	事業担当部署-担当部署2	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
47	事業担当部署-担当者役職2	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
48	事業担当部署-氏名2	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
49	製造・輸入時期2	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
50	購入・販売時期2	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
51	設置期間2	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
52	使用期間2	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
53	事業者-事業者名3	0	0.00	0.0	16	0.00	20.8	0	0.00	0.0
54	事業者-事業区分3	0	0.00	0.0	6	0.00	16.1	0	0.00	0.0
55	事業担当部署-所在地3	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
56	事業担当部署-電話番号3	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
57	事業担当部署-担当部署3	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
58	事業担当部署-担当者役職3	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
59	事業担当部署-氏名3	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
60	製造・輸入時期3	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
61	購入・販売時期3	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
62	設置期間3	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
63	使用期間3	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
64	措置状況詳細	4,679	1.00	253.8	1,282	0.23	185.2	1,014	0.74	186.0
65	原因究明等-原因調査状況	4	1.00	6.8	4	0.34	6.3	3	0.88	6.1
66	原因究明等-事故原因	16	1.00	25.2	29	0.27	36.3	9	0.06	45.9
67	原因究明等-事故原因詳細	9,816	1.00	349.2	1,462	0.32	130.6	5,616	0.90	476.3
68	原因調査機関-機関名	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
69	原因調査機関-連絡先	0	0.00	0.0	0	0.00	0.0	9	0.00	28.4
70	通報者情報-氏名	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
71	通報者情報-属性	0	0.00	0.0	733	1.00	24.7	0	0.00	0.0
72	通報者情報-職業	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
73	通報者情報-住所	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
74	通報者情報-連絡先-TEL	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
75	通報者情報-連絡先-FAX	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
76	通報者情報-連絡先-E-MAIL	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
77	最終更新年月日	31	1.00	16.9	667	1.00	16.7	217	1.00	16.8
78	受付年月日	2,135	1.00	16.9	1,844	1.00	16.7	1,917	1.00	16.7
79	情報元システムID	20,060	1.00	9.9	17,653	1.00	10.9	7,901	1.00	10.0
80	情報提供機関名、担当課名-所属	18	1.00	16.2	3	0.99	12.2	1	1.00	15.0
81	情報提供機関名、担当課名-氏名	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0
82	情報提供機関名、担当課名-TEL	0	0.00	0.0	0	0.00	0.0	0	0.00	0.0

次に、上記で得られた情報に基づき、人手により分類項目をタイプ別に分類した。図表２－10に、この分類作業のために設定したタイプ種別をまとめる。タイプの分類にあたっては、辞書サイズおよび平均バイト数を参照しながら、具体的な登録情報内容を目視で確認した。ただし、個々の情報提供元における登録環境の詳細（たとえばレコード登録時に選択リストが登録者に提示されるか等）は考慮していないため、現在の運用環境の実態と本調査の分類は必ずしも対応しないことに注意が必要である。

【図表２－10】分析に用いた登録項目の種別

種別	説明	分類項目の例
I	識別番号	《事故情報ID》など
P	個人情報で分析対象としないもの	《通報者情報-氏名》など
L	場所	《発生場所》など
D	時間・時期	《発生日時》など
T0	リスト選択で辞書が固定されているもの（データベース運用期間を通して語彙が変化しないもの）	《事故種別》など
T	リスト選択で対象が固有名であるもの（データベース運用期間中に語彙を追加する可能性があるもの）	《商品など名称》など
TS	リスト選択に近いが一般名詞であるもの	《傷病内容詳細》など
S	自由記述文	《事故内容詳細》など

（イ）言語解析の適用可能性に関する調査

（ア）の調査結果に基づき、《情報提供元》が《国民生活センター》で、《事故種別》が《事故情報》である85,778レコードを抽出して言語解析の適用可能性を調査した。対象となるレコード群には、自由記述項目の平均バイト数が約890バイトと、言語解析の適用に十分な分量のテキストが含まれている。また、《事故種別》を《事故情報》に限定したことから、《傷病内容》や《傷病の程度》の網羅率が1.0となり、本調査のテキストマイニングの目的に合致したサンプルとなっている。調査では、さらに、網羅率等を踏まえて、元データに含まれる83個の分類項目から合計17個を解析対象として選んだ。ここで、《発生年》と《発生月》は元の事故情報データでは１つの分類項目として《発生日時》にまとめられていたが、季節やイベントと関連がある事故を分析するため、異なる分類項目として再定義した。また、商品に関する情報、事業者に関する情報はそれぞれ、《商品情報》および《事業者情報》としてまとめた。以上により、合計で14個の分析用分類項目を設定した。

解析にあたっては、これら14個の分類項目をさらに、「絞り込み項目」と「表示項目」の２つに振り分けた。「絞り込み項目」とは、図表２－10のL（場所）、D（時間・時期）、T0（リスト選択）であり、値を指定することで条件に合致するレコードを絞り込むことができる。「表示項目」とは、T（固有名）、TS（一般名詞）、S（自由記述文）であり、辞書サイズが大きいことから絞り込みには適さないと判断した。分析に用いたデータの概要を図表２－11に示す。

【図表２－11】言語解析を適用した85,778レコード集合に関する統計データ

項目種別	項目分類	分類項目（オリジナル）	分類項目（分析用）	辞書サイズ	網羅率	平均バイト数
D	絞り込み項目	発生日時	発生年	46	0.57	4.0
D			発生月	12	0.53	2.0
L		発生場所-施設用途	発生場所-施設用途	9	0.70	13.4
L		発生場所-発生場所	発生場所-発生場所	10	0.10	7.7
T0		被害者の情報-所属-職業	被害者の情報-所属-職業	10	1.00	13.0
T0		被害者の情報-年代等-年代	被害者の情報-年代等-年代	11	0.91	8.2
T0		被害者の情報-性別	被害者の情報-性別	3	0.99	6.0
T0		事故内容	事故内容	17	1.00	11.1
T0		傷病内容	傷病内容	20	1.00	19.6
T0		傷病の程度	傷病の程度	7	1.00	15.4
TS	表示項目	事故概要	事故概要	23,564	1.00	17.1
T		商品など名称	商品情報	32,485	1.00	24.6
T		型式・ロット・生産国		19,051	0.31	31.8
T		事業者-事業者名1	事業者情報	31,033	0.82	18.3
T		事業者-事業区分1		1,034	0.02	44.1
T		事業者-事業者名2		6,196	0.16	17.3
T		事業者-事業区分2		58	0.00	39.1
S		事故内容詳細	事故内容詳細	85,761	1.00	889.8

本調査における分析は、(A)絞り込み項目を使ったレコードのグループ化、(B)グループごとの項目要約表示、の２つのステップから構成される。

(ステップA) 絞り込み項目を使ったレコードのグループ化

図表２－11で絞り込み項目として指定した10個の項目の任意の組み合わせを指定して、これらの項目について値が同じレコードどうしを

グループ化する。たとえば、分類項目《発生月》には《１月》から《12月》の値が入力されていて辞書サイズは12、分類項目《傷病の程度》には《医者にかからず》、《治療１週間未満》、《１～２週間》、《３週間～１か月》、《１か月以上》、《死亡》、《不明》のいずれかが入力されていて辞書サイズは７である。たとえば、絞り込み項目として、《発生月》と《傷病の程度》の２つを指定する場合には、「《４月》に発生した《治療１週間未満の事故》」など、組み合わせ数合計12×７＝84個のグループが得られることになる。この方法では、任意個の絞り込み項目を自由に組み合わせて、機械的に多数のグループを生成することができる。一方で、生成されるグループの数が膨大になり、人手で確認するべきグループを見つけることが難しくなるという問題がある。そこで、得られたグループの中から注目するべきグループを選別するため、当該グループに含まれる事故レコード数と意外性を表す統計尺度（相互情報量）を掛け合わせた尺度を使って、グループの順位づけを行った。

(ステップB)グループごとの項目要約表示

（ステップA)で得られた各グループの項目ごとに、指定した数の特徴的な「キーワード」を表示する。これによって、各グループの特徴を項目別に概観することが可能になる。ここでキーワードとは、《事故内容詳細》以外の項目については入力されている文字列そのもの、《事故内容詳細》については自由記述文に対して言語解析（形態素解析と係り受け解析）を適用した結果から得られる名詞句、および名詞句と動詞の係り受けペアとする。さらに、抽出した名詞句や係り受けペアが各グループにおいてどれくらい特徴的であるかを統計的な尺度（頻度×相互情報量）を用いて計算し、そのスコアに基づき上位のものを出力した。

エ　分析の試行を通して明らかになった点

子どもの重大事故情報に、上記の分析技術を適用した例を以下に示す。

まず、上記で作成したサンプルデータで《傷病の程度》が《１か月以上》または《死亡》であるレコードの中から、《被害者の情報-所属-職業》および《被害者の情報-年代等-年代》を手がかりに、中学生以下と判断できるものを抽出して、《０歳以下》、《１～４歳》、《５歳以上未就学児》、《小学生》、《中学生》のラベルを付与した。登録レコードの中には、「40歳代の小学生」なども存在するため、不適切と思われるレコードの

一部は人手で除いた。

全体で、《死亡》事故９件、《１か月以上》の傷病にかかわる事故487件が抽出された。判断の目安とした値の組み合わせ、および、抽出されたレコード数を図表２－12にまとめる。

【図表２－12】子ども事故分析のためのラベル付与

《被害者の情報-所属-職業》	《被害者の情報-年代等-年代》	ラベル	事故発生件数
無職	1～4歳	1～4歳	124
小学生	5～9歳	小学生	95
中学生	10歳代	中学生	80
小学生	10歳代	小学生	45
無職	0歳以下	0歳以下	44
無職	5～9歳	5歳以上未就学児	42
その他	1～4歳	1～4歳	15
その他	5～9歳	5歳以上未就学児	13
その他, 学生	5～9歳	5歳以上未就学児	12
小学生		小学生	8
その他	0歳以下	0歳以下	7
その他, 不明	5～9歳	5歳以上未就学児	3
その他, 不明	1～4歳	1～4歳	3
その他, 学生	1～4歳	1～4歳	3
中学生		中学生	1
小学生	1～4歳	小学生	1

本調査では、絞り込み項目の数を２個から３個に設定して、合計4,459件のグループを出力した。図表２－13に最も事故発生件数が高かったグループの表示例を示す。「||」は区切り記号、括弧内は発生件数である。

また、特徴的なグループの上位３件は、①《中学生》の《化粧石鹸》による《皮膚障害》、②《住宅》における《ウォータークーラー》による《熱傷》、③《住宅》における《化学物質による危険》による《呼吸器障害》であった。これらのグループの詳細は、図表２－13の形式の要約を参照したり、人手でグループ内の事故情報を個別に調べたりすることで、さらに確認することが可能である。

【図表２－13】本調査における分析手法で得られた事故グループの例

事故発生件数	59
絞り込み項目	《発生場所・施設用途》店舗・商業施設　《事故内容》その他　《傷病内容》骨折
発生年	2013(15)，2016(9)，2012(8)，2014(6)，2008(1)，2010(6)，2015(3)，2009(1)，2011(4)
発生月	01(6)，08(5)，04(5)，12(4)，02(5)，05(8)，03(4)，10(3)，07(4)，09(3)
発生場所-施設用途	（絞り込み項目として設定）
発生場所-発生場所	その他(4)，階段(1)，玄関(1)
被害者の情報-所属-職業	その他，不明(2)，小学生(19)，その他，学生(3)，その他(5)，中学生(7)，無職(23)
被害者の情報-年代等-年代	1～4歳(21)，5～9歳(11)，小学生(19)，中学生(7)，0歳以下(1)
被害者の情報-性別	男性(37)，不明(2)，女性(20)
事故内容	（絞り込み項目として設定）
傷病内容	（絞り込み項目として設定）
傷病の程度	1か月以上(59)
事故概要	遊具(3)，屋内遊戯施設(2)，施設内遊園地のエアー滑り台(2)，スポーツ教室(2)，遊園地(2)，鍼灸院のマッサージ(1)，他の建物(1)，カラオケボックス(1)，商業施設内の遊戯施設の遊具(1)，遊具施設(1)，…
商品情報	遊具(3)，施設内遊園地のエアー滑り台-滑り台-遊園地・レジャーランド(2)，○○○○(2)，店のキッズコーナーエアマット-遊具-店舗.事務所(1)，有料ボールプール-遊具-遊興施設利用(1)，有料遊具施設-遊具(1)，住居雑品その他(1)，他の建物-他の保育サービス(1)，バレエ教室(1)，昇降設備(1)，…
事業者情報	○○○○
事故内容詳細（名詞句）	遊具(30)，子供(67)，ボール プール(12)，責任(23)，息子(65)，救急 車(26)，無料(16)，店 側(12)，ギブス(10)，腕(15)，
事故内容詳細（係り受け）	全治(21)，怪我|を|する(18)，骨折(7)，左 ひじ|を|骨折 する(5)，不満(10)，ボール プール|に|飛び込む(4)，子供|を|遊ば せ て いる(4)，娘|を|遊ば せ て いる(4)，責任|は|ない(5)，損害 賠償|を|求める(5)，子供|が|骨折 する(4)，責任|を|認める(4)，条例|は|ある(3)，一切 責任|を|負う(3)，…

最後にまとめを述べる。重大事故の分析は人手で詳細に行う必要があるが、その前段階として分析対象を絞り込み整理することが有効である。その具体的な方策を検討する際に、本調査で報告した手法が参考になると考えられる。

オ　分析の試行を通して課題と感じられた点

事故情報データバンクでは、複数の情報提供元からの事故情報を収集して、横断的に検索する仕組みを提供している。このようなサービスは事故情報の一元的な分析の第一歩となるものであるが、一方で、情報提

供元により「辞書サイズ」、「網羅率」、「平均バイト数」の傾向が大きく異なることは、マイニングを実施する際に注意を要する。異なる情報提供元どうしで事故情報を共有するだけではなく、事故情報の入力インタフェースや整合性チェックツールを共有することが、データの品質向上に結び付くと期待される。

事故情報データバンク中では、「子ども」、「高齢者」、「外国人」など、分析の中で重要な役割を果たす概念が明示的に定義されていない場合がある。このため図表２－12で試みたように、分析に先立ってデータの整理や不整合なデータへの目視による対応などが必要になる。テキストマイニングの目的に応じて辞書を整備することが必須であると考えられる。

本調査では、事故情報データベース中の分類項目を、「絞り込み項目」と「表示項目」の２種類に振り分けて分析を試みた。絞り込み項目で値が入力されていない欠損値について、自由記述項目のテキストを利用して情報を補完する仕組みを検討することも今後の検討課題である。

（４）事故情報データ分析（市瀬 龍太郎）

ア　分析目的

事故情報データバンクでは、大量の事故情報を保持している。起きる事故には、一定のパターンが存在する。そのようなパターンを発見することで、頻繁に起こる事故の特徴を抽出することを目的とする。

イ　分析手法

本分析においては、頻繁に起こる事故の特徴を抽出するために、データマイニングにおいてしばしば利用される頻出パターンマイニング手法を応用する。頻出パターンマイニングとは、事例同士で頻繁に共起する事象を抽出する手法であり、スーパーにおいて精算時の買い物かごの中に入っている商品の組み合わせから、購買傾向を把握することなどに使われる。例えば、ソーセージを買う人の多くがロールパンを購入しているといった傾向を抽出することができ、マーケティングなどにも応用される技術である。本分析においては、事故情報データで頻繁に共起する語を利用することで、起こる事故のパターンの抽出を試みる。

本分析では、２段階の手続きを取り、事故の特徴の抽出を試みた。最初は、事故のグループ化である。そのために、対象の事故データの「事故概要」、「商品など分類」から名詞を抽出し、抽出された名詞が頻繁に共起する事例をグループとして抽出した。次に、各グループ内の特徴を抽出するために、グループ内の事故の「事故内容詳細」から名詞を抽出し、抽出された名詞で頻繁に共起する名詞をグループ内の事故を特徴づける名詞として列挙した。

ウ　対象データ

事故は、被害の重大さによって異なる傾向を持っていることが予想される。また、子供に起こる事故と、高齢者に起こる事故も異なる特徴を持っていることが予想される。そのため、被害者の重傷度と年齢により、いくつかのカテゴリに分けて分析を行った（図表２－14）。図は縦軸に被害者の重傷度、横軸に被害者の年齢をとったものである。重傷度は、重いものから死亡、重傷（治療に１か月以上）、その他に分類されており、その中から、「死亡」「重傷・死亡」の場合に着目した。年齢は、９歳以下、10～59歳、60歳代、70歳代、80歳以上で分類を行い、子供と高齢者に着目した。重傷度と年齢の組み合わせにより、図表２－15に掲載した14個のカテゴリを設定し分析を行った。なお、データの総件数は、

198,135件である。

【図表２－14】重傷度と年齢による分類

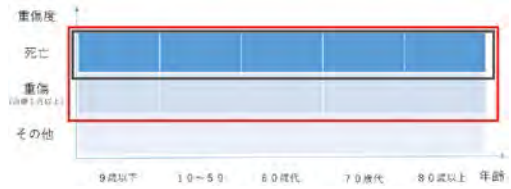


図１：重傷度による分類

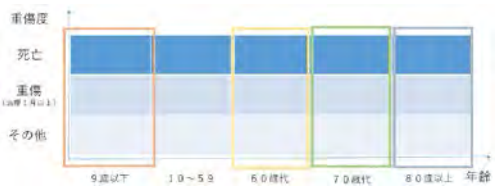


図２：年齢による分類

【図表２－15】年齢と重傷度の組み合わせによる14個の分析カテゴリ

1.	重傷・死亡	13,621件	8.	６０歳代，死亡	53件
2.	死亡	1,517件	9.	７０歳代（全て）	8,972件
3.	９歳以下（全て）	3,889件	10.	７０歳代，重傷・死亡	1,541件
4.	９歳以下，重傷・死亡	555件	11.	７０歳代，死亡	97件
5.	９歳以下，死亡	74件	12.	８０歳以上（全て）	4,366件
6.	６０歳代（全て）	13,191件	13.	８０歳以上，重傷・死亡	1,037件
7.	６０歳代，重傷・死亡	1,916件	14.	８０歳以上，死亡	208件

エ　分析結果

（ア）カテゴリ１（重傷・死亡）：抽出グループ数122個

５語以上の名詞によりグループ化されたものは30個。

グループ例：「洗顔、石鹸、化粧、衛生――顔、皮膚、受診」

「住宅、設備、建物、構成、材――救急、転倒、骨折」

「住居、生活、用品――製品、建物、火災」

ここで、グループ例の最初に挙げた語は、グループを作成する際に使われた「事故概要」「商品など分類」に見られた頻出語の例であり、後に挙げた語は、グループの特徴をみるための「事故内容詳細」の頻出語の例である。

このカテゴリでは、医療、保健関連の事例などが見られる。

【図表２－16】カテゴリ１（重傷・死亡）

ID	出現数	語1	語2	語3	語4	語5	語6	語7
24	492	洗顔	衛生	保健	品	石鹸	類	化粧
41	612	機器	医療	保健	福祉	サービス	衛生	品
64	886	理	美容	保健	福祉	サービス	衛生	品
9	448	器具	理	保健	用品	美容	衛生	品
46	618	機器	医療	保健	衛生	品	サービス	
65	887	理	美容	保健	衛生	品	サービス	
66	896	理	美容	保健	福祉	サービス	品	
42	613	機器	福祉	保健	サービス	衛生	品	
71	916	美容	福祉	保健	サービス	衛生	品	
78	1012	医療	保健	福祉	サービス	衛生	品	
34	534	洗顔	衛生	保健	品	化粧	類	
90	1231	石鹸	衛生	保健	品	化粧	類	
4	435	美容	医療	保健	福祉	サービス		
43	613	機器	医療	保健	福祉	サービス		
47	619	機器	衛生	品	保健	サービス		
72	917	美容	衛生	品	保健	サービス		
79	1018	医療	保健	衛生	品	サービス		
87	1168	理	美容	保健	福祉	サービス		
115	2883	化粧	類	品	保健	衛生		
28	504	構成	材	住宅	設備	建物		
10	449	器具	理	保健	用品	美容		
7	447	用品	福祉	保健	サービス	品		
11	450	理	美容	保健	用品	品		
59	846	機器	医療	保健	衛生	品		
67	897	理	美容	保健	サービス	品		
68	904	住	生活	住居	用品	品		
74	926	美容	福祉	保健	サービス	品		
76	962	理	美容	保健	衛生	品		
80	1049	医療	保健	福祉	サービス	品		
100	1629	福祉	サービス	衛生	品	保健		

（注）　ID　　：カテゴリ内グループの識別子
　　出現数　　：語が含まれる事故の数（語は順不同）

（イ）カテゴリ２（死亡）：抽出グループ数102個[20]

６語以上の名詞によりグループ化されたものは24個。

グループ例：「ベッド、介護、生活、用品、住居――すき間、ボード、首」

「体育、指導、課外、部、活動――（体操、練習）（野球、バッティング）（サッカー、部、活動）」

「医薬品、医療、サービス――投与、過剰、中毒」

「住居、開放、式、石油、ストーブ――火災、全焼」

表記方法は、カテゴリ１と同様であるが、２番目の例では、グループ内で体操、野球、サッカーといくつかの代表的なケースが見られたので分けて記述した。

[20] （イ）以降の図表については、第31回消費者安全専門調査会　市瀬消費者安全専門調査会専門委員資料参照。
http://www.cao.go.jp/consumer/kabusoshiki/anzen/doc/031_170612_shiryou5.pdf

このカテゴリでは、火災、介護、部活動、医薬品などの事例が見られる。

（ウ）カテゴリ３（９歳以下）：抽出グループ数112個
５語以上の名詞によりグループ化されたものは25個。
グループ例：「建物、設備、サービス――（カビ、アパート、部屋）（スイミング、スクール）」
「娯楽、遊具――（滑り台、骨折）（ショッピング、センター）」
「外食、サービス――アナフィラキシー、メニュー」
このカテゴリでは、アレルギーや遊んでいる最中などの事例が見られる。

（エ）カテゴリ４（９歳以下、重傷・死亡）：抽出グループ数60個
４語以上の名詞によりグループ化されたものは27個。
グループ例：「建物、設備、構成、材――自動、ドア、怪我」
「乗り物、車両――車、指」
「娯楽、遊具――テーマパーク、滑り台」
このカテゴリでは、乗り物、遊んでいる最中などの事例が見られる。

（オ）カテゴリ５（９歳以下、死亡）：抽出グループ数24個
４語以上の名詞によりグループ化されたものは９個。
グループ例：「体育、教科――（運動、場）（プール、水泳）」
「認可、外、保育、施設――就寝、幼児」
このカテゴリでは、学校活動、保育中などの事例が見られる。

（カ）カテゴリ６（60歳代）：抽出グループ数99個
４語以上の名詞によりグループ化されたものは33個。
グループ例：「住宅、設備――骨折、転倒、治療」
「調理、食品――冷凍、ピザ、農薬」
「美容、サービス――毛、染め、皮膚」
このカテゴリでは、医療、保健、サービスなどに関する事例が見られる。

（キ）カテゴリ７（60歳代、重傷・死亡）：抽出グループ数60個
４語以上の名詞によりグループ化されたものは28個。

グループ例：「住宅、設備――（シックハウス、頭痛）（転倒、骨折）」
「医療、サービス――（歯医者、抜歯）（視力、矯正）」
このカテゴリでは、医療、保健のサービス、住宅に関する事例などが見られる。

（ク）カテゴリ８（60歳代、死亡）：抽出グループ数37個
４語以上の名詞によりグループ化されたものは23個。
グループ例：「住居、生活、用品、手すり――ベッド、すき間」
「食料、食品――のど、窒息」
このカテゴリでは、福祉、医療などの事例が見られる。

（ケ）カテゴリ９（70歳代）：抽出グループ数65個
４語以上の名詞によりグループ化されたものは24個。
グループ例：「医療、サービス――イン、プラント」
「石鹸、化粧――皮膚、科、小麦」
このカテゴリでは、医療、保健、サービスに関する事例などが見られる。

（コ）カテゴリ10（70歳代、重傷・死亡）：抽出グループ数62個
５語以上の名詞によりグループ化されたものは17個。
グループ例：「保健、福祉、医療、機器――（イン、プラント、歯科）（採血、注射、しびれ）（金属、アレルギー）」
「乗合、バス、サービス――バス停、停車、転倒、骨折」
このカテゴリでは、医療、保健のサービス、バスに関する事例などが見られる。

（サ）カテゴリ11（70歳代、死亡）：抽出グループ数52個
５語以上の名詞によりグループ化されたものは20個。
グループ例：「食品、嗜好、品――（団子、喉）（パン、喉）」
「電動、車いす、移動――（転落、水田）（踏切、列車）」
「医療、過誤――透析、入院」
このカテゴリでは、介護、医療、車いすなどに関する事例が見られる。

（シ）カテゴリ12（80歳以上）：抽出グループ数95個
５語以上の名詞によりグループ化されたものは19個。

グループ例：「住居、生活、用品、サービス――介護、大腿、骨」
「医療、機器、サービス――入れ歯、歯科」
「乗合、バス、サービス――バス停、発車、転倒、骨折」
このカテゴリでは、医療、保健、サービス、バスなどに関する事例が見られる。

（ス）カテゴリ13（80歳以上、重傷・死亡）：抽出グループ数62個
４語以上の名詞によりグループ化されたものは27個。
グループ例：「乗合、バス、運輸――バス停、発車、転倒、骨折」
「建物、設備――（スポーツ、クラブ、階段）（スーパー、床、水）」
このカテゴリでは、バスや福祉などに関する事例が見られる。

（セ）カテゴリ14（80歳以上、死亡）：抽出グループ数48個
４語以上の名詞によりグループ化されたものは23個。
グループ例：「医療、機器――電位、治療」
「養護、老人、ホーム――ご飯、のど」
「乗り物、移動――電動、車いす」
このカテゴリでは、介護、保健などに関する事例が見られる。

オ　分析における課題・留意点

分析の際に一番大きな課題となったのは、記入方法、入力用語が統一されていない点である。日付など、入力の際に標準化ができるものは、なるべく標準化することで、機械的な分析により、類似事例の把握などが容易になると考えられる。また、発生場所など未入力のデータも多く、統一的な傾向の分析が行えない属性も見られた。分析の際に、どのような分析をするかを視野に入れながら、必要に応じて必須入力事項などの設定をすることも考えられる。

本分析においては、カテゴリ内の事故に対して、いくつかのグループに分類を行った。しかし、語の共起頻度に基づく分類のため、グループ内のすべての事故が同じ形態とは限らず、グループ内においても複数のパターンが存在する可能性がある。また、同一の事故が複数のグループに分類される場合もあるため、似たグループの多さが同様の事例の多さを意味するものではない。このように分析手法により、固有の特性があるため、データ分析の際には、それを理解した上で、目的に応じた分析手法を設計する必要がある。データ分析は、全体の傾向を掴み、さらに

深い分析を行うための糸口であり、詳細な分析には、人間の目による個別事故の確認が必要である点にも留意が必要である。

（５）テキストマイニングを用いた事故データ分析の試行と考察（西田 佳史）

ア　はじめに

事故対策や傷害予防で効果的な方法の一つは、関連する製品を特定し、その改善策を開発することである。近年、子どもや高齢者の製品に関連した事故が多発しており、データを活用した製品改善の方法論の開発が求められていることから、こうした生活機能変化者（心身機能や認知機能が変化しやすい子どもや高齢者）や製品関連事故を今回の分析の対象とする。

本調査の目的は、事故情報データバンクに登録されている事故情報に関して、テキストマイニング技術等を活用した分析を行うことにより、事故対策（事故に起因する傷害の予防）を行うべき対象の明確化や、事故対策の効果評価に対してデータを活用する可能性を検討する。

本稿では、事故状況が記載された自由記述文を分析することで、子どもと高齢者が関与した重症事故と軽症事故の比較分析することで、重要領域の選定の可能性の検討をおこなう。また、今回、当委員会を通じて入手したデータ以外に、関連するデータベースの分析事例として、日本スポーツ振興センターの災害共済給付データを活用したトレンド分析（介入の効果評価）の可能性についても関連分析のケーススタディとして示す。

以下、試行結果とその考察を述べる。

イ　テキストマイニングを用いた分析結果

○　子どもの事故と高齢者の事故の分析

事故情報データバンクに登録されている、子どもと高齢者の事故データについて分析を行った。

今回は、「事故内容詳細」の項目について、テキストマイニングを行い、名詞（製品名）を抽出した。その上で、各製品名が含まれているインシデントの数（全体数）と、そのうち、重傷１名、死亡１名、又は重症１名以上が含まれているインシデント数（重症以上数）をカウントし、その比率（重症以上数／全体数）を算出した。

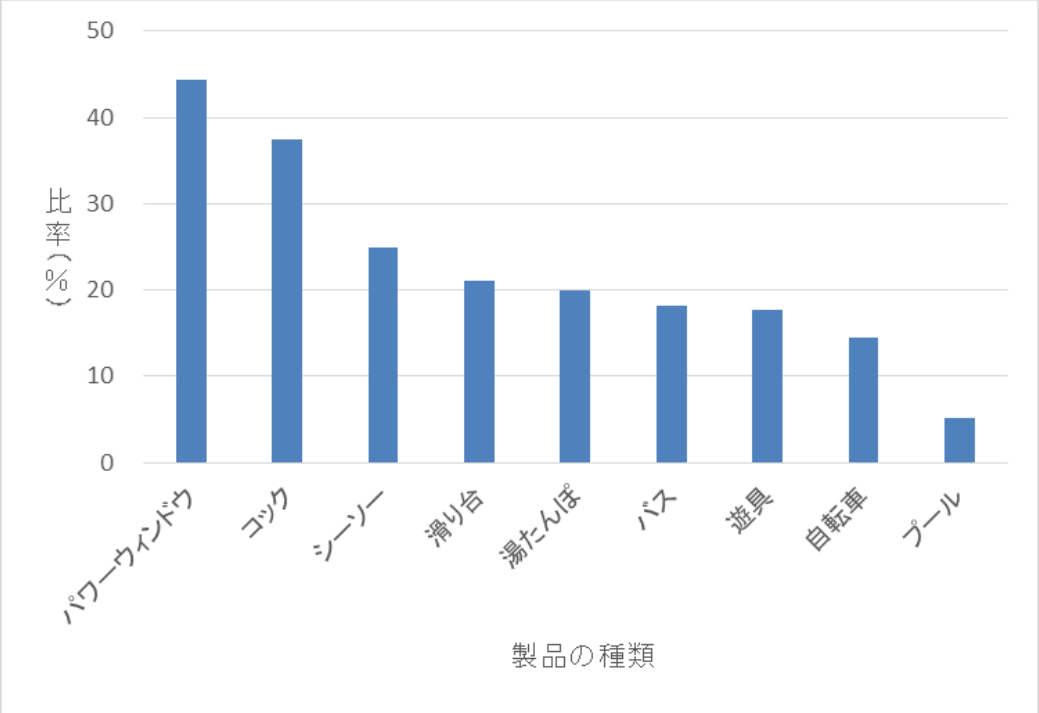
子どもについては、事故データの項目の「被害者の情報-年代等-年代」が「０歳以下」、「１～４歳」、「５～９歳」における全事故に出現する製品名に対して、重症以上数で出現する製品名の比率を算出した。

高齢者に関しては、「被害者の情報-年代等-年代」が「60歳代」、「70歳代」、「80歳以上」における全事故に出現する製品名に対して、重症

以上数で出現する製品名の比率を算出した。

図表２－17に、９歳以下の分析結果を示した。図中「コック」は、「ウォーターサーバー」のコックである。縦軸の比率は、上述したとおり、重傷１名、死亡１名、又は重症１名以上である事故を「重症以上」とし、これが全体に占める割合を意味している。上位の製品は、パワーウィンドウ、コック、シーソー、滑り台、湯たんぽ、バス、遊具、自転車、プールなどとなった。

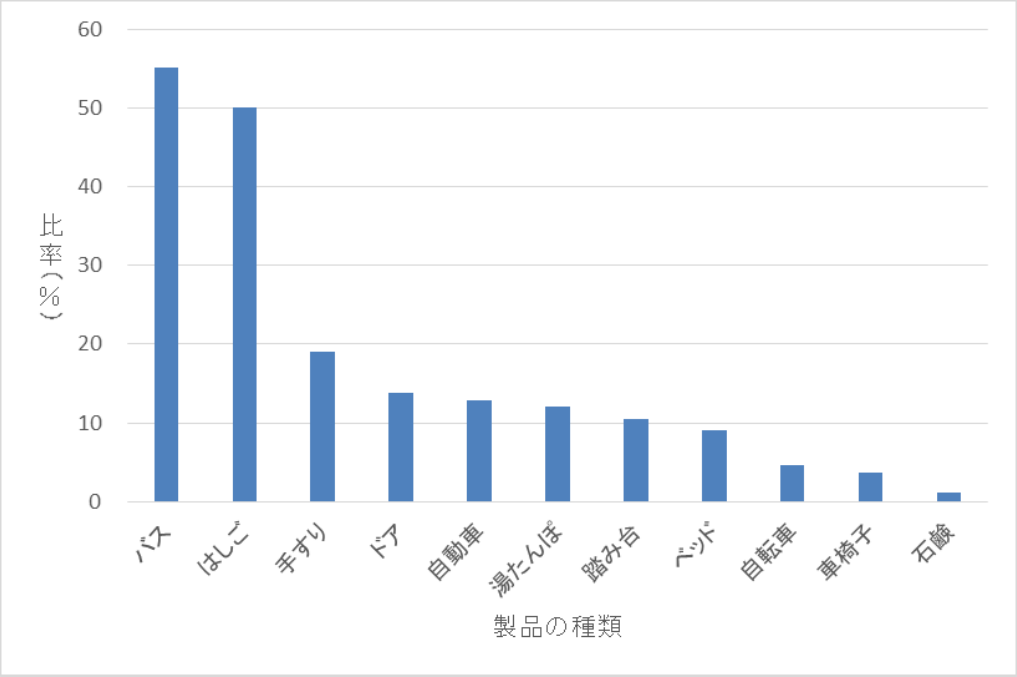
【図表２－17】９歳以下の事故データの分析結果



製品名	全体数（件）	重症以上数（件）	比率（%）
パワーウィンドウ	9	4	44.4
コック	8	3	37.5
シーソー	8	2	25.0
滑り台	38	8	21.1
湯たんぽ	10	2	20.0
バス	11	2	18.2
遊具	96	17	17.7
自転車	103	15	14.6
プール	39	2	5.1

図表２－18に、60歳以上の分析結果を示した。上位の製品は、バス、はしご、手すり、ドア、自動車、湯たんぽ、踏み台、ベッド、自転車などとなった。

【図表２－18】60歳以上の事故データ分析結果



製品名	全体数（件）	重症以上数（件）	比率（%）
バス	388	214	55.2
はしご	8	4	50.0
手すり	58	11	19.0
ドア	203	28	13.8
自動車	62	8	12.9
湯たんぽ	33	4	12.1
踏み台	19	2	10.5
ベッド	277	25	9.0
自転車	149	7	4.7
車椅子	164	6	3.7
石鹸	1,498	17	1.1

これらの結果からテキストマイニング技術を用いることで、事故に関連した製品のリストの作成や、各製品と重症事故の関係を分析可能であることが分かった。これらの技術は、基本的には、インシデント報告書の中に含まれる製品名を数え上げているため、事故の直接的原

因が製品にあったのか、そうでなかったのか、単に書かれていただけの無関係な製品なのかまでを自動で行うことは現時点でも課題があると考えられる。そのため、重要トピックの候補の作成後は、人の目による精査も必要である。しかしながら、膨大なデータに対して、これを行うことは困難であるため、解析者がこうした人工知能を活用して作業を進めることは有用であると考えられる。適用と限界に関しては、「エ　考察」で改めて議論したい。

ウ　他のデータを用いたスポーツ外傷（柔道）のトレンド分析

学校管理下で発生した事故による傷害を科学的に予防するためには、対策や予防につながる重要な情報を含んでいる自由記述文を分析することが求められているが、現状では、膨大な自由記述文の処理は解析者の手作業に頼っていることから、詳細な分析は困難である。本研究では、人工知能技術（テキストマイング）を援用することで、介入の効果評価を行う技術の実現可能性を検討する。

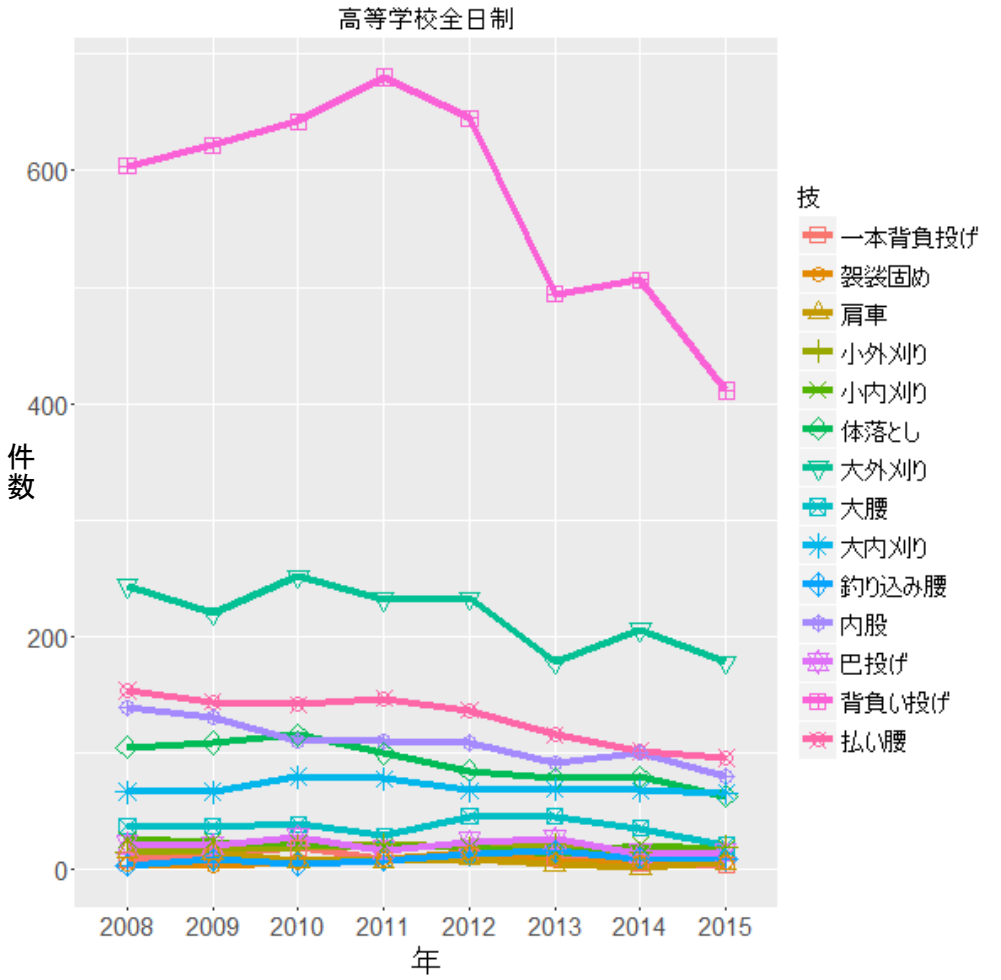
日本スポーツ振興センターの保有する災害共済給付データのうち柔道における災害事故ビッグデータ152,695件のうち被災年が2008年から2015年の８年間における高等学校の事故データ60,300件に対して、テキストマイングを適用することで経年変化分析（トレンド分析）を行った。傷害の種類の経年変化分析に加え、テキストマイングを使わなければ困難な分析として、柔道の技に着目し、傷害に関連した柔道の技の経年変化を分析した。

図表２－19に分析結果を示した。傷害に関連した柔道の技の経年変化の分析から、「背負い投げ」が、2013年から顕著に減少していた。2008年と比較すると、2015年は34%減少していた。「大外刈り」も減少していたが、2008年と比較すると、2015年は17%減少であった。柔道では、2013年に事故に関する書籍が発行されるなど、社会問題化し、注意喚起が行われた。特に、「背負い投げ」と「大外刈り」のリスクが指摘された。経年変化分析から、こうした注意喚起の効果が、「背負い投げ」と「大外刈り」に関して確認された。

以上の分析により、テキストマイニング技術を用いることで、あらかじめ項目立てされていない項目を事後的に設定し、その数を調査したり、数の経年変化を調べることが可能であることが分かった。事後的な設定に関しては、事故報告書のフォーマットの修正などの際に、新たな項目の候補を作るなどの応用が考えられる。また、数の経年変化の分析に関しては、注意喚起や安全基準作成などの何らかの対策・介入を行った際

に、それがどのような効果に繋がったのかを分析することに繋がる機能であり、効果評価への応用が考えられる。

【図表２－19】柔道の技と事故の関係の分析



エ　考察（テキストマイニング技術の適用可能性と限界に関する留意点）

今回、テキストマイング技術をインシデントデータに適用することで、どのような分析が可能かの検討をおこなった。消費者事故データの適用結果からは、事故に関連した製品の抽出や、製品ごとに重症事故が占める割合などの分析が可能であり、重要なトピックの候補を抽出することに有効であると考えられる。基本的には、文章に含まれる単語を数え上げるため、厳密には、本当に事故を起こした製品なのか、単に書かれていただけかを自動で判断することは、今なお課題があると考えられるが、トレンド分析による効果評価など、トレンドを把握したい、もしくは、他の種別とのインシデントと相対比較をしてみたい、という応用に関し

ては、有効であると考えられる。

一方、基本的には、単語の数え上げなので、対象として取り上げたいインシデント以外のものが含まれてしまっている可能性がある。そのため、裁判のための資料など絶対数が重要な応用には適切ではない場合がある。少なくとも限界を示したうえで数値を使うことが求められる。

以上より、テキストマイニングのあるべき使用方法としては、以下のものが考えらえる。重要トピックのフィルタリング、注目すべきトピックの候補リスト作成、頻度の経年変化を見ることによる効果評価、事故を記録する用紙やフォーマットの変更のための新たな項目の候補リストの作成、などである。

２　事業者による分析

ここでは、株式会社プラスアルファ・コンサルティング[21]による、事故情報データバンクに集約されたデータの検討・分析結果を整理する[22]。

（１）テキストマイニングによる事故内容詳細（自由記述）の分析

テキストマイニングの手法を用いることにより、自由記述に含まれている情報を、どのような形で把握することが可能か、検討・分析を行った。

ア　単語のランキング

まず、自由記述を単語レベルに分解し、名詞、動詞それぞれを、出現頻度別にランキングした（図表２－20）。

名詞ランキングを示すことにより、事故、トラブルとなる対象を把握することができる。

また、動詞ランキングを示すことにより、事故、トラブルの原因となる消費者の行動を把握することができる。

【図表２－20】テキストマイニングにより抽出した名詞及び動詞ランキング

名詞ランキング

No.	単語	品詞	件数	割合
1	火災	名詞	1,840	6.6%
2	走行中	名詞	1,293	4.6%
3	対応	名詞	1,289	4.6%
4	メーカー	名詞	1,247	4.5%
5	業者	名詞	920	3.3%
6	発煙	名詞	918	3.3%
7	エンジン	名詞	892	3.2%
8	ネット通販	名詞	730	2.6%
9	情報	名詞	698	2.5%
10	治療費	名詞	684	2.5%

動詞ランキング

No.	単語	品詞	件数	割合
1	買う	動詞	4,888	17.5%
2	焼損する	動詞	2,948	10.6%
3	出る	動詞	2,495	9.0%
4	発生する	動詞	2,225	8.0%
5	使う	動詞	1,778	6.4%
6	出火する	動詞	1,498	5.4%
7	する	動詞	1,424	5.1%
8	含める	動詞	1,177	4.2%
9	食べる	動詞	1,109	4.0%
10	求める	動詞	950	3.4%

[21] http://www.pa-consul.co.jp/corporate/index.html

[22] テキストマイニングによる分析の対象は、「事故情報データバンク」で公開されているデータのうち、2015 年 11 月 1 日から 2016 年 10 月 31 日の間のデータ（登録日ベース、27,820 件）とした。

イ　単語のランキングから係り受けの関係をマップ化

主語と述語の関係（係り受け）を認識させることにより、事故の内容をより詳細に把握することが可能である（図表２－21）。

ここでは、「買う」を中心において、購入製品、購入場所、その他の関連事項をつなげた。図表２－21には、「SNS広告」の「ネット通販」で「お試し」の「健康食品」を「購入」したら３回分の契約になってしまったといった、消費者トラブルが含まれている。

この係り受けをランキングすることにより、単語ランキングよりも、詳細に、どのような事故が多く発生しているのかを把握できる可能性がある。

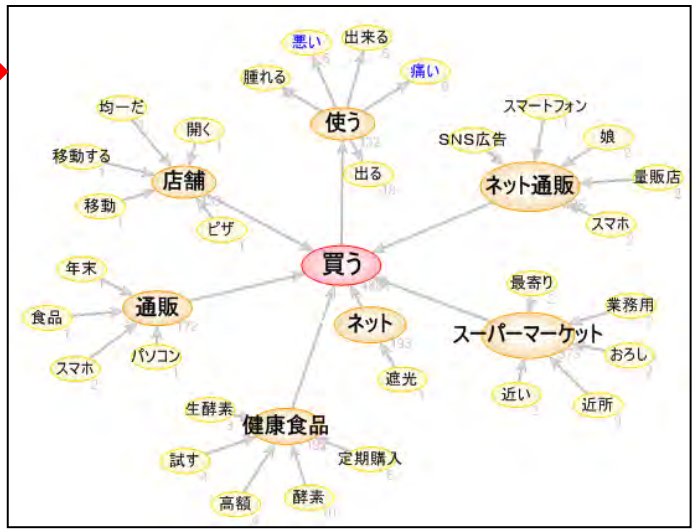
例えば、この時期の事故情報を分析すると、「パソコン-内部-燃焼する」「タブレット端末用ACアダプター-コード部分-出る」等の係り受けが上位にあった。

【図表２－21】動詞から係り受けの関係をマップ化

動詞ランキング

No.	単語	品詞	件数
1	買う	動詞	4,888
2	焼損する	動詞	2,948
3	出る	動詞	2,495
4	発生する	動詞	2,225
5	使う	動詞	1,778
6	出火する	動詞	1,498
7	する	動詞	1,424
8	含める	動詞	1,177
9	食べる	動詞	1,109
10	求める	動詞	950

係り受け



ウ　共起の関係にある単語同士を線で結ぶことにより情報を分類

同じ文中に、一緒に発言される（共起の関係）単語同士をつなぐことにより、事故情報別のマップを作成することが可能である（図表２－22）。

この期間、子どもが足、指に火傷を負う、化粧品で肌トラブル、顔が腫れるといった問題等が生じていたことが分かる。