Research Report of the National Institute
of Industrial Safety, NIIS-RR-2002(2003)
UDC 614.8:681.3

爆発火災災害データベースにおける用語間の関係付けに関する研究*

大塚輝人**，板垣晴彦**

Study of Terms Connection in the Database for Explosion and Fire Accidents*

by Teruhito Otsuka**, Haruhiko Itagaki**

***Abstract*:** According to the development of internet, various search methods have been invented recently. Because of large number of web pages, page ranking sysytem has been succeeded on internet. But such method is not applicable for the database for explosion and fire accidents, because of lack of linkage system and paucity of same accident.

For the database, there is another requirement, the outputs should be not only the direct reference by the words and phrases which have completely in agreement but also the guessed data which was able to set in order with a certain aspect such as similarity. From the guessed data, the point to notice of the proccess can be found previously.

In this study, a simple but quite powerful technique, quantification method of third type, is used for deriving quantitative information from qualitative data, which made from abstracts of accidents. The several way for getting keywords from abstracts in Japanese are compared. As a result, the quantitive distance relation among terms was obtained reasonably.

The merits of introduction of the quantitive distance relation among terms are summurized as below;

1) Since it is quantitive information, placing in the order is possible to outputs.
2) The correlation of the noun obtained can be used not only in the database for explosion and fire accidents used for distance space creation, but in another database with a perspective of explosion and fire accidents.
3) It is possible to adjust by selection of the axis, the scale of a specific axis, and the applicable number of cases etc. into a parameter for the reduction of the information noise.
4) According to the simplified nouns, it is possible to take partial match to a reference word. Therefore, reference by the compound noun which does not apear in the database can also return some information.

***Keywords*;** Relation between terms, Explosion and fire accidents, Information analysis, Quantification method of the third type, Cluster analysis

1. はじめに

従来，データベースの取り扱いは，あらかじめデータ登録者によって登録されたキーワード群，ならびに分類番号によって階層化され，そのインデックスを追うことによって目的データにたどり着くようになっている。階層化にともなう分類を適切に行うことはデータ登録時の大きな負担であり，かつ，分類境界上に存在するデータの処理は登録者の主観に任されている。このような分類をテキストデータの分解により自動的かつ客観的に行うための研究もなされている[2), 3), 4)]。

* 平成14年12月 6 日，第35回安全工学研究発表会[1)]において一部発表
** 化学安全研究グループ Chemical Safety Research Group

近年におけるコンピュータの高速化が全文検索を可能にしたが，検索者が目的データにたどり着くためにはデータ登録者と同じ語彙の中から適切な検索語を選ばねばならない状況は変わっていない。このため，検索者は幾度もの試行錯誤を繰り返さねばならず，その試行錯誤も報われるとは限らない。この点において，その試行錯誤という面倒な行為を軽減し，完全に一致する語句による直接の検索のみではなく，類推されるデータの順序付けられた出力が求められている。データの類似度が定量化されれば，前記のような登録者側の階層化や分類の手間も軽減される。

災害データベースにおいては同一の災害というものは基本的に起こることはないため，試行錯誤を繰り返さねばならないという傾向は顕著である。また，データの作成者が必ずしも対象災害を専門としない場合に適切な用語を用いること，あるいは類似と判別すべき災害が時間的あるいは空間的に離れて起きていた場合に同一の作成者がデータを作成したとしても表現の一貫性を保つことは，非常に困難である。**Table 1** に同種の災害防止対策が有効と考えられるが，起きた現象の記述には同じ名詞がほとんど用いられていない例を示す。実際の作業現場や，災害調査では，現象の記述を基に類似災害を検索できるデータベースが求められている。

化学物質を取り扱う作業現場での災害では，さらに事情は複雑である。化学物質の名称としてIUPAC（国際純正および応用化学連合）の規則および勧告と，CAS（ケミカルアブストラクツサービス）とが存在するが，それら物質の正式名が作業現場で使われることはまずない。通常，資料段階では化学式あるいは一般名等が記載され，さらに実際に作業を行う作業者たちにより略称化されている。したがって，一次調査を行う者がどの段階の名称を採取するかに関しての統一はほとんど不可能である。

以上のような問題点の解決の第一歩として，データベース上のキーワード間の関係を数量化し距離付けすることで，キーワード間の客観的かつ定量的な比較を行えるようにすることが，本研究の目的である。

2. 手法

2.1 データベース

本研究に用いたデータベースは，国内の爆発火災災害（1955年から）を対象とし本研究所化学安全研究グループによって独自に作成された非公開のものである。詳細な内容については文献[5]に詳しい。本研究に供したのは，研究開始時に入力が完了していた1996年までの41年分5513件中，災害の概要部分の記載がある5066件の自由記述された文章データである。概要の文章中には主に火災，爆発発生の経緯が記されている。また，原因がわかっている場合にはその原因についても言及されている。**Table 1** の例 2 はその一例である。

Table 1　Example of accidents in similar situation. 類似災害の例（例 1 は安全工学vol.41No.5（2002）の自己災害ニュース，例 2 は爆発火災災害データベースから）

例 1:血圧降下剤の製造中に酢酸エチルの入った反応がまにアミノ酸化合物を入れたところ爆発し、最上部の投入口から火が噴出した。酢酸エチルは危険物第 4 類第 1 石油類に指定されており，静電気で引火した可能性がある。
例 2:製薬工場の展着室において，缶入りの薬品が固くて開かないので暖房機のそばに置き２０分くらい温めた。そして塗布作業を開始したところ，引火爆発してストーブ周辺にいた７名が火傷を負い，火災により建物を焼失した。原因は，揮発したベンゾール蒸気が室内に充満し，静電気火花か何らかの点火源により引火爆発したとみられる。

2.2 形態素解析

通常の検索行動から考えて，対象を名詞に絞ることとした。その名詞を自由記述された文章から抽出するための道具として，形態素登録が比較的簡易である京都大学で開発されたJUMANを用いた。JUMANの形態素辞書に載っていない語，例えば「電極」，「気鐘」，「爆ごう」，「突沸」などの名詞と，解析を正確にするための接頭辞，接尾辞，動詞など合わせて約400語を追加登録してから，解析を行った。解析例を**Table 2** に示す。通常の文法と異なり，「揮発する」のような動詞も，サ変名詞＋動詞として「揮発・する」に分解されている。このようなサ変名詞も今回の解析に含めた。

2.3 形態素解析の後処理

JUMANにおいては，外来語をカタカナ化してそのまま用いるという日本語の柔軟性から，形態素辞書の発散を避ける意味でカタカナ語は解析されず，未定義語として処理される。そのため，未定義語を名詞に含め，さらに，数詞および形式名詞を排除した。その結果延べ名詞数12万 4 千語9095種の名詞が得られた。得

Table 2 Example of morphological analysis.
形態素解析例

揮発	きはつ	揮発	サ変名詞
した	した	する	動詞
ベンゾール	ベンゾール	ベンゾール	カタカナ
蒸気	じょうき	蒸気	普通名詞
が	が	が	格助詞
室内	しつない	室内	普通名詞
に	に	に	格助詞
充満	じゅうまん	充満	サ変名詞
し	し	する	動詞
,	,	,	読点
静電気	せいでんき	静電気	普通名詞
火花	ひばな	火花	普通名詞
か	か	か	接続助詞
何らか	なんらか	何らか	副詞
の	の	の	接続助詞
点火	てんか	点火	サ変名詞
源	げん	源	名詞性名詞接尾辞
に	に	に	格助詞
より	より	よる	動詞
引火	いんか	引火	サ変名詞
爆発	ばくはつ	爆発	サ変名詞
した	した	する	動詞
と	と	と	格助詞
み	み	みる	動詞
られる	られる	られる	動詞性接尾辞
。	。	。	句点

られた名詞の中には，JUMANの辞書に含まれた多くの複合語が一つの名詞として帰属されているものがあるため，二通りの後処理を行った。そのうち一つは，連続する接頭辞，名詞，接尾辞をまとめてひとつの名詞として扱った全複合名詞化処理（以下「複合化」という）であり，もう一つは得られた名詞を可能な限り分解した全単一名詞化処理（以下「単一化」という）である。当然ながら，複合化では名詞の種類数は増えるが延べ名詞数は減少し，単一化では名詞の種類数は減少し，延べ名詞数は増える。

単一化では，カタカナ語の表記，送り仮名の表記，漢字の表記の統一を行う必要があったため，同時に，アルファベットによる略号の化学物質名化，化学物質名の官能基名分割，同義語の統一を行った。帰属できなかったアルファベットや記号による略号は，この時点で取り除いた。以上の作業で作成されたデータの特徴を一覧表の形で，**Table 3** に示す。

Table 3 には，単一化処理の後，同一災害にのみ現れる名詞群を一つの単語としてカテゴライズした同一視化処理（以下「同一視化」という）を行ったデータもあわせて提示した。数量化Ⅲ類の計算量が（サンプル数とカテゴリー数の小さい方）^{2}×（サンプル数とカテゴリー数の大きい方）に比例するため，同一視化によって計算量の削減が可能である。

本研究においては，災害間の比較により名詞の関連性を得ることが目的であるため，同一災害内での同一名詞の繰り返しは数えず，0，1のチェックシートとしてコード化した。

高頻度の名詞について，複合化では「引火爆発」が一語として出てきており，「爆発」の頻度は他に比べて低い。また，単一化では「着火」と「引火」を同義語と見なしているため，大きく頻度があがっている。単一化と同一視化において延べ語数と高頻度の名詞が同一となるのは，低頻度の名詞が主にグループ化されるためである。

低頻度の名詞では，表現のゆれ（ex.「洗浄」と「洗滌」など）や明らかに分割が必要なもの（ex.「酸素アセチレン炎」）が，複合名詞には含まれる。JUMANの辞書に含まれる単語の基準が正確な形態素解析を目指すものであるため，多くの表現，複合語が含まれている。同一視化においては，同一視した組を挙げてある。そこには意味的なつながりはなく，共起性が100%であるにすぎない。同一視化を行わずに単一化のまま数量化Ⅲ類を行った場合でも，それらの名詞は同じスコア（カテゴリーウェイト）を持つ。

3. 数量化Ⅲ類の結果

前節のデータを数量化Ⅲ類に供して得られた固有値と相関係数を**Table 4** に示す。延べ名詞数，あるいは，名詞種類数の多い方が相関係数は大きくなっている。これはあるサンプル（災害事例）にのみ現れるカテゴリー（名詞）が多ければ多いほど，（サンプルウェイト）=（カテゴリーウェイト）の軸近くに多くの点をプロットすることが出来るようになるためである。また，複合化を推し進めていけば，最終的に事例の概要全体を一カテゴリーとして扱うような例を考えることが出来るが，この場合サンプル数＝カテゴリー数と成り，$y=x$となる任意の取り方が相関係数1を与える。このことからも，前述の相関係数が良くなることを説明できる。

複合化と，単一化の名詞に関する数量化による散布

Table 3　Post processes after morphological analysis of the database for explosion and fire accidents and their feature.
爆発火災災害データベースの形態素解析後の処理とその特徴

処理名	複合化	無処理	単一化	同一視化
主な処理内容	連続した名詞を一語にし，前後の接頭辞，接尾辞を名詞に接続	形式名詞等の排除	複合名詞を分解し，表記，同義語の統一	単一化の処理に加えて同一災害にのみ出てくる語を同一視
高頻度の名詞(カッコ内は該当件数)	爆発 (2802)	爆発 (3222)	爆発 (3412)	爆発 (3412)
	原因 (1831)	原因 (1879)	稼働 (2398)	稼働 (2398)
	引火 (761)	作業 (1834)	着火 (2212)	着火 (2212)
	火傷 (754)	引火 (1253)	原因 (1879)	原因 (1879)
	火災 (653)	ガス (971)	器 (1508)	器 (1508)
	火 (561)	工場 (858)	ガス (1413)	ガス (1413)
	被災 (544)	火傷 (814)	火炎 (953)	火炎 (953)
	ガス (534)	被災 (751)	工場 (867)	工場 (867)
	引火爆発 (518)	着火 (713)	火傷 (819)	火傷 (819)
	発生 (489)	火災 (701)	周り (815)	周り (815)
低頻度の名詞(該当件数 1)	アルカリ洗い工程	アルキルフェノール	不均一	かしわ, 精肉, 選任, 養豚
	アルカリ洗浄	アルキルフェノールプラント	無菌	まっこう, 鯨油
	アルカリ洗滌後	アルミダイカスト	亜酸化窒素	ウェイトレス, バーテン
	アルカリ洗滌塔	アルミダイカストマシン	亜硝酸	カルボン酸, ゼリー, 転移
	隧道建設工事	アルミダイキャスト	硼酸	クロレラ, 園児, 汁粉, 乳糖, 保育
	隧道建設工事現場	酸素アセチレン炎	蔗糖	ゲリラ, 過激派, 時限
	隧道建設作業	人絹	蛆虫	延縄, 設け, 鮪
	隧道工事	薬きょう	饅頭	鰹節, 防虫
	隧道工事現場	薬莢	麩	雇用, 軟鉄
	他計 17040 語	他計 4313 語	他計 2145 語	他計 1550 語
延べ該当件数	84223	102660	112116	110939
延べ名詞数	91551	123663	139971	139971
名詞種類数	23897	9095	5800	5189

Table 4　Eigen values of quantification method of third type.
数量化III数の結果の固有値（カッコ内は相関関数）

複合化	無処理	単一化	同一視化
0.7669(0.8757)	0.6510(0.8068)	0.4692(0.6850)	0.4409(0.6640)
0.7456(0.8635)	0.5120(0.7156)	0.4446(0.6668)	0.3965(0.6297)
0.7370(0.8585)	0.5014(0.7081)	0.4226(0.6501)	0.3892(0.6239)
0.7236(0.8506)	0.5006(0.7075)	0.3973(0.6303)	0.3463(0.5885)
0.7171(0.8468)	0.4888(0.6991)	0.3756(0.6129)	0.3368(0.5803)

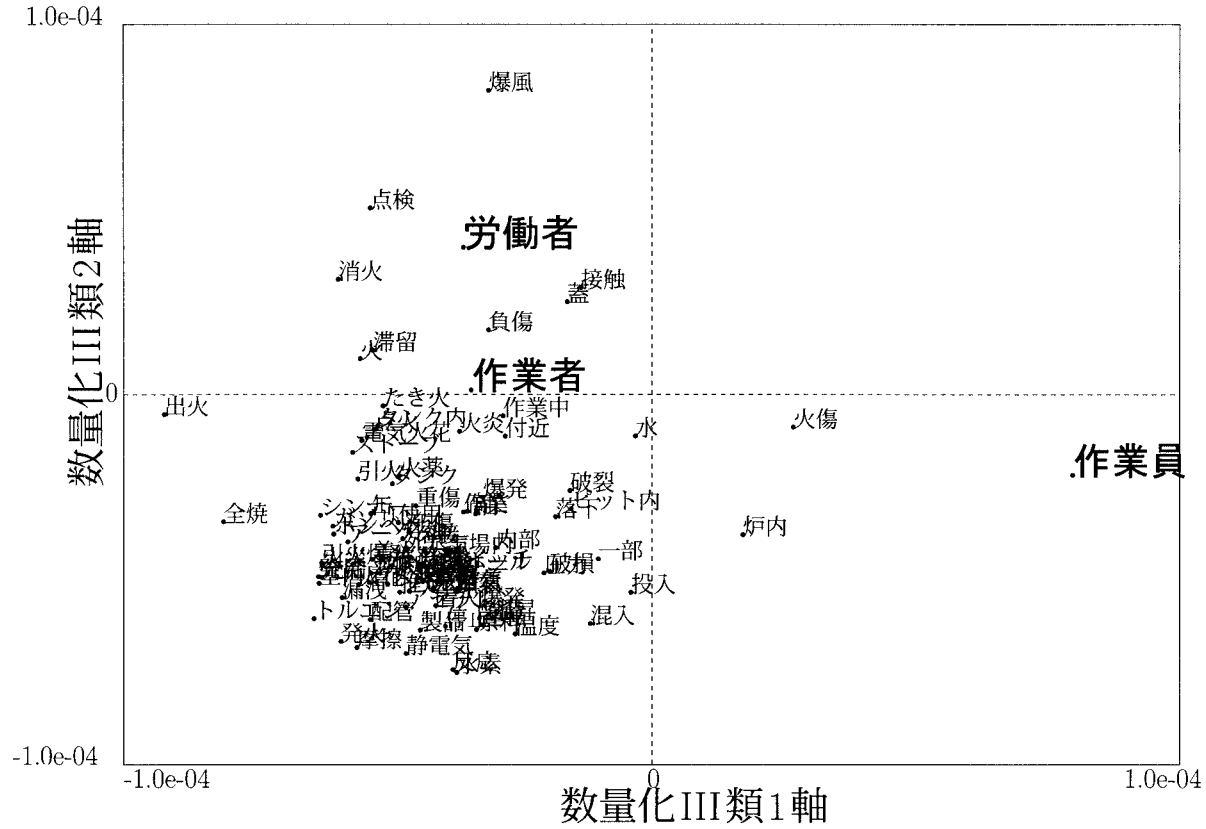

Fig. 1 Scatter diagram of compound noun by the quantification method of the third type.
複合化によって作成された数量化III類の１，２軸による散布図（本文中で言及されている名詞をボールド体で示した）

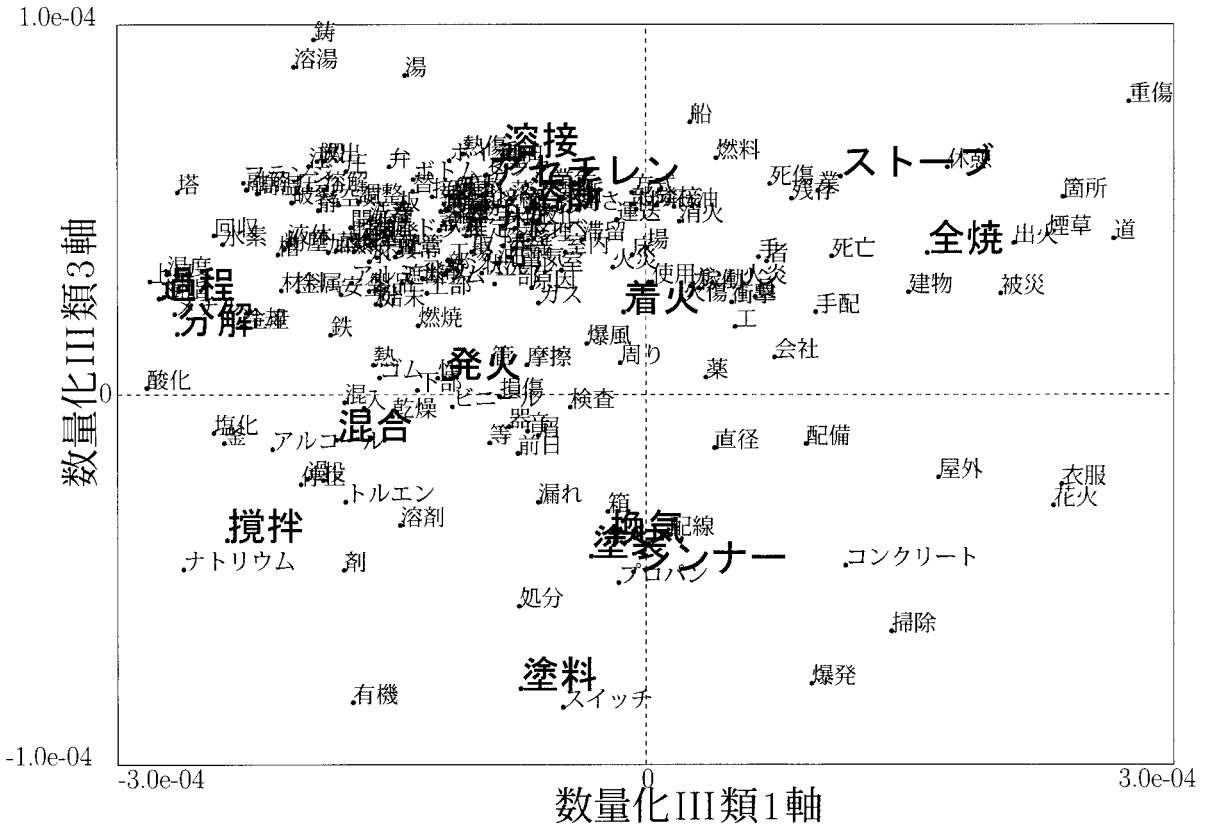

Fig. 2 Scatter diagram of single noun by the quantification method of the third type.
単一化によって作成された数量化III類の１，３軸による散布図（本文中で言及されている名詞をボールド体で示した）

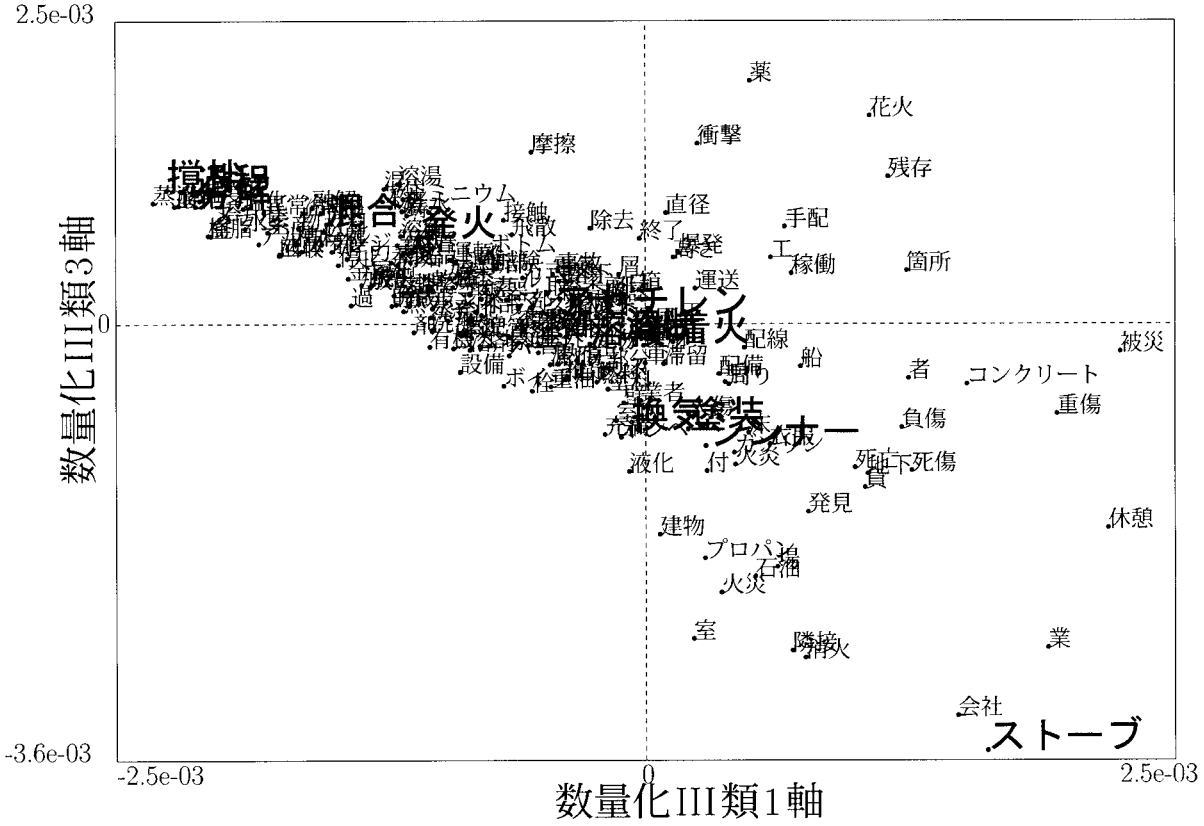

Fig. 3　Scatter diagram of unified data by the quantification method of the third type.
同一視化によって作成された数量化Ⅲ類の１，３軸による散布図（本文中で言及されている名詞をボールド体で示した）

図の一部を**Fig. 1** と**Fig. 2** とに示す。散布図は，該当件数100件以上の名詞をプロットした中の約９割である複合化104語中97語，単一化230語中208語を含むような部分を拡大したものである。

複合化の結果では，同義語と考えてよい「作業員」「作業者」「労働者」が散在してしまい，軸に意味を見出すことが難しくなっている。そのため，本研究の目的である概念間の距離付けには不向きであると言える。

それに対して単一化の場合，１軸マイナス方向に反応プロセス関係の言葉，「過程」，「分解」，「混合」，「撹拌」等が，プラス方向に火災に関する言葉，「ストーブ」「全焼」等がグループを形作っている。範囲外でプロットされていないが，「ダイナマイト」，「発破」，「トンネル」等も１軸についてプラスの大きな値を取っている。また，「発火」が比較的プロセスよりであるのに対して「着火」は火災よりであり，両者とも中央の比較的近い位置にあることも注目に値する。

２番目の固有値に属する固有ベクトルから得られた分布は，１軸とほぼ同傾向（１軸との相関係数0.9437）を示したため，ここではy軸として３軸を用いている。３軸方向では「塗装」，「換気」，「シンナー」などがマイナス方向でグループを作っており，「溶接」，「溶断」とそれに用いる「アセチレン」などがプラス方向でグループを形作っている。

Fig. 3 に同一視化による数量化Ⅲ類の１軸と３軸を用いた散布図の一部を示す。図には**Fig. 1** と**Fig. 2** と同様に230語中約９割の209語が含まれている。１軸によるプロセスと火災とのグループ化は，単一化の場合を良く再現しているが，y軸方向に割り当てた３軸では，単一化で分離されていた名詞が分離されておらず，単一化の良い近似とは見なされない。また，この傾向は２軸でも同様であった。これは，同一視化によって固有値（相関係数の二乗値）が下がり，２，３軸が説明軸として取るには不十分なものになってしまったためではないかと考えられる。逆に，単一化においても固有値0.3965が一つの目安であると考えることが出来る。

以上のことから，名詞間の関係を明らかにするためには単一化が望ましく，同一視化による相関係数の低減は爆発火災災害データベースにおいては無視できないものであるといえる。

Fig. 4 に示したのは，単一化したものを数量化した結果の１-４軸を用いて最遠隣法を用いて得られたデンドログラム（樹形図）である。ここでは該当件数に

ついて，上位80語という高頻度な名詞のみについて作成した。距離の閾値を適当に設定することで，グループ化された各々のクラスターを得ることも可能である。

このようにして作成された名詞の各クラスターは，元となった災害データベースと関係なく，関連語として用いることができ，かつ，ヒットしたデータ間について，得点付けが可能となる。また，クラスター化を経ずとも，用語間の距離計算を直接行って，与えられた名詞に対して距離の短い，関連度の高い名詞を順序付けることも可能である。

4. 関連語抽出への利用

単一化したデータを用いた場合の実際の関連語抽出への手順は以下のようになる。

1）入力

入力は単語一語でも，複数でも構わない。また，形態素解析を利用するので，文章によって入力してもよい。

2）単語の単一名詞化

データベースを単一化した手順に従い，形態素解析と単一化フィルタを利用することにより単一名詞を得る。入力から得られた名詞が，データベース内の単一化された名詞にマッチしない場合，すなわち元々データベース内に存在しなかった名詞についてもN-gram化して部分一致を行うことで，入力の有意な部分を拾い出すことも可能である。

3）名詞間距離の計算

得られた単一化された名詞からの距離（関連度）を計算する。出力する最長距離（最低関連度）をあらかじめ決め，メッシュで散布図を区切っておくことにより全名詞についての冗長な計算を避けることが可能である。

4）得られた名詞の出力

得られた名詞を距離の近い順に表示する。ソートの計算量を減らす意味でも最長距離を既定しておくことが望ましい。

以上の手続きを行った例を**Table 5** に示す。距離の計算には 1 - 4 軸についての二乗和の平方根を用い，該当件数100以上のものについて上位10語を表にしたものである。ここでも火災と生産過程のグループが現れており，また，「マグネシウム」のような物質名についても類似災害を引き起こしやすい「アルミニウム」が高位に現れている。通常の類語辞典，あるいは関連語情報との決定的な違いは，定量的な関連度が定義できる点にある。この関連度情報を用いれば，データベースの該当データを類似度順にソートするなど，より有意な情報を上位に並べることができる。

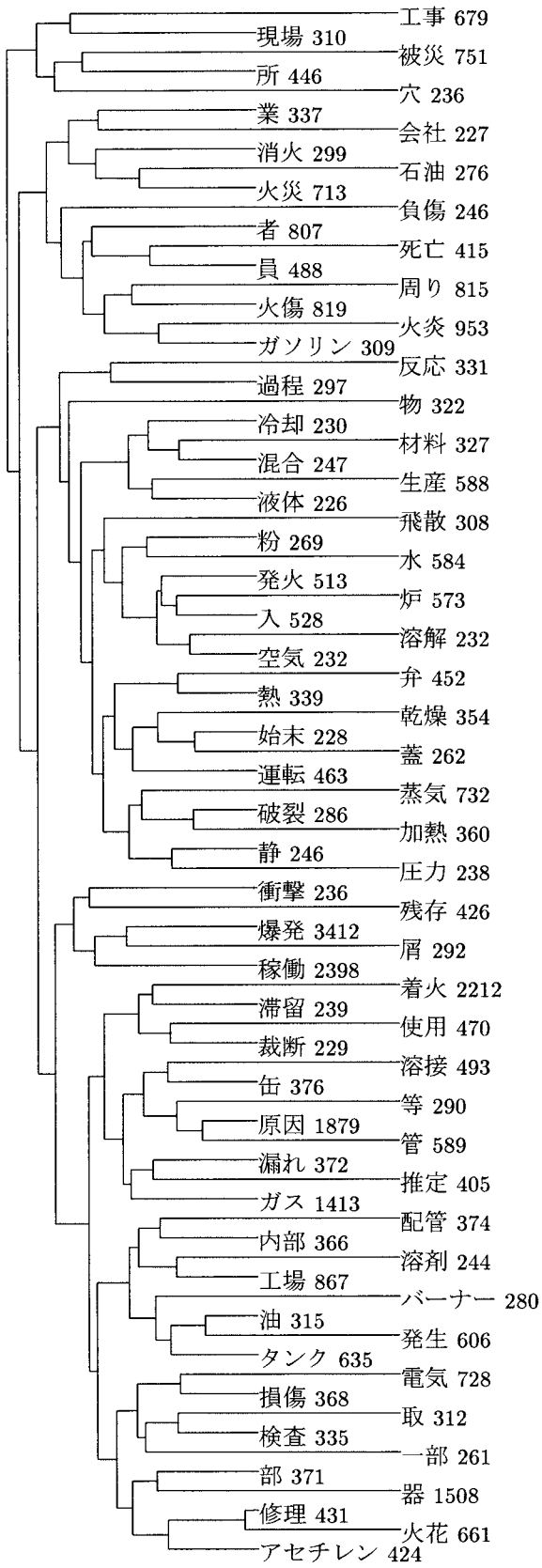

Fig. 4 Dendrogram by cluster analysis.
最遠隣法を用いたクラスター分析による樹形図

Table 5　Distance ordered related nouns.
距離（関連度）順の関連名詞

キーワード	火災		マグネシウム		生産過程 (生産・過程)	
順位	名詞	距離	名詞	距離	名詞	距離
1	石油	2.2964e-04	投	1.8984e-04	温度	6.5877e-05
2	室	2.7687e-04	アルミニウム	1.9250e-04	分解	9.1195e-05
3	場	3.5066e-04	高温	1.9839e-04	アルコール	9.5275e-05
4	プロパン	3.8490e-04	融解	1.9960e-04	水素	1.0547e-04
5	隣接	4.6302e-04	混	2.0014e-04	酸化	1.0856e-04
6	消火	6.3674e-04	粉	2.4338e-04	上昇	1.3052e-04
7	建物	6.9386e-04	溶湯	2.4618e-04	槽	1.4604e-04
8	発見	6.9767e-04	混合	2.5768e-04	撹拌	1.6128e-04
9	付	1.0714e-03	冷却	2.7330e-04	液体	1.6734e-04
10	火炎	1.0741e-03	袋	2.8459e-04	異常	1.8255e-04

5. 結言

本研究では，データベース検索の際のキーワードの観点から「該当・非該当」ではない，関連度による順位付けられた出力に向けての第一歩として，災害データベースにおける名詞の定量的な関連付けを行った。その結果は見てきたとおり良好なものであった。また，本研究で得られた定量的な関連付けを行った名詞群には以下のような利点がある。

1）定量的な情報であるため，出力に順位付けが可能である。
2）距離空間作成に利用した爆発火災災害データベースに限らず，別のデータベースでも得られた名詞の相互関係は，爆発火災災害に着目して検索する場合に利用が可能である。
3）情報量としてのノイズ低減のために，用いる軸の選択，特定軸のスケール，該当件数によるフィルタリング等をパラメータとして，利用時に調整することが可能である。
4）名詞が単一化されているので，検索語に対して部分一致を取ることが可能である。したがってデータベースに含まれていない複合語による検索でも，何らかの情報を返すことが可能である。

参考文献

1）大塚輝人，板垣晴彦，爆発火災災害データベースにおける用語間の関係付けに関する研究，第35回安全工学研究発表会講演予稿集，2002.
2）鈴木芳美，建設工事労働災害に関するテキスト情報の解析，産業安全研究所研究報告RIIS-RR-92，pp.103-105，1993.
3）鈴木芳美，建設工事労働災害事例の発生状況記録中のフリータームの統計解析，産業安全研究所研究報告RIIS-RR-93，pp.89-95，1994.
4）鈴木芳美，建設工事労働災害の発生状況の記録における情報構造に関する多変量統計解析，産業安全研究所研究報告NIIS-RR-96，pp.11-22，1997.
5）爆発・火災災害の統計分析，産業安全研究所安全資料NIIS-SD-No.15，1997.
6）黒橋禎夫，長尾真，日本語形態素解析システムJUMANver.3.61マニュアル，1999.
7）林知己夫，駒澤勉，数量化理論とデータ処理，朝倉書店，1982.
8）脇本和昌，後藤昌司，松原義弘，多変量グラフ解析法，朝倉書店，1979.

（平成15年 2 月 7 日受理）