目的に応じた複数意味空間連結機能を有する法学分野文書対象意味検索システム

藤川 滉† 佐々木 史織‡ 清木 康†

†慶應義塾大学 環境情報学部 〒252-8520 神奈川県藤沢市遠藤 5322
‡慶應義塾大学 政策・メディア研究科 〒252-8520 神奈川県藤沢市遠藤 5322
E-mail: †{t09828af, kiyoki}@sfc.keio.ac.jp, ‡sashiori@sfc.keio.ac.jp

あらまし 本稿では，複数の意味空間の連結により，利用者の目的に応じた類似文書の検索を行う意味検索システムを提案する．本システムの特徴は，法学分野に関する複数の意味空間を構成し，予め定義された専門用語の分類リストによって，文書群の各文書中に含まれる専門用語を類似の意味を持つグループに集約し，TF-IDF 等の指標により計算された各文書における単語の重要度を表す値を用いて，検索目的に応じた意味空間上にマッピングすることにより，指定分野に着目した検索や全体的な意味内容を計量する統合的な検索を実現する複数意味空間連結機能である．本システムは，ユーザが指定した文書群中の任意の文書を対象として，文書群の各文書との相関度を計算し，検索結果を生成する．本システムの利用者は，目的に応じた効率的な検索をすることが可能になる．

キーワード 意味空間の連結，意味検索システム，TF-IDF，テキストデータベース

1. はじめに

コンピュータが大衆化するにつれ，電子化された文書に触れる機会が多くなっており，またユーザ自らも文書を電子データとして蓄積する傾向にある．情報の蓄積はインターネット上にもおよび，そこには膨大な情報群が存在している．

ところが，文書を検索する手段としては，旧態依然としたキーワードによる検索に頼らざるを得ないのが現状である．

インターネット上には電子化された判例等の多種多様な法律関連文書が存在しているが，キーワード検索のみによる文書検索では，これらの膨大かつ複雑な意味的関連性を持つ情報群のなかから利用者の目的とする法律関連知識を体系的かつ適切に発見することは難しい．

キーワードによる検索を行う一般的な検索エンジンにおいて，間接的・意味的に関連する文書を効率的に獲得することは難しい．法律に特化した専門のデータベース[1][2]においても，キーワード検索が主であり，参照条文や，判例を参照している別の判例を検索する機能などは存在するものの，間接的・意味的に関連する文書を効率的に獲得する手段としては十分ではない．

従来のキーワード検索では，直接キーワードを指定する必要があり，指定したキーワードを含まないものの，間接的に関連性の高い文書や，利用者の意図・目的と意味的に関連性の高い文書を検索することは難しい．

例えば，ある法律学に関する文書を持っており，その内容と類似度の高い文書を検索したいという要求があったとしても，文書に含まれるキーワードを直接指定する必要があるほか，法律学の特定の分野に関する文書を検索するなどの，利用者の意図・目的に沿った検索を行い，適切な検索結果を得ることは困難である．

本稿では，検索の目的によってその目的に適した意味空間上に文書をマッピングし，それら複数の意味空間の連結により，利用者の目的に応じた，より利用者の意図に近い類似文書の検索を行う意味検索システムを提案する．

本システムの特徴は，法律関連の複数の意味空間を構成し，予め定義された専門用語の分類リストによって，文書群の各文書中に含まれる専門用語を類似の意味を持つグループに集約し，各文書中の単語について計算された「単語重要度」（TF-IDF 等の指標により計算された，各文書における単語の重要度を表す値）を用いて，検索目的に応じた意味空間上に文書群中の各文書をマッピングし，「特定分野のみに着目した検索」や「全体的な意味内容を計量する統合的な検索」を選択することを実現する「複数意味空間連結機能」にある．

本システムは，利用者が指定した文書群中の任意の文書を対象として，文書群の各文書との相関度を計算し，検索結果を生成する．本システムの利用者は，検索目的に応じた効率的な類似文書の検索が可能となる．また，本システムの基本方式は，法律分野以外の大テーマ，小テーマを持つような分野に関する文書を対象として，目的に応じたマッピング対象空間を変更することにより，類似する文書を的確に検索することができ，多様な分野の文書分析に応用可能である．

2. 基本方式

本システムは，複数の意味空間の連結により，利用者の目的に応じた類似文書の検索を行う意味検索システムである．

本システムは，利用者の指定する文書に対して相関度計算を行い，相関度の高い文書を計量して出力する．

2.1. 意味空間の形成

法律分野の文書を対象とした意味空間の形成についての基本アイデアを図 1 に示す．本稿では，一例として，検索対象を法律学の一分野である憲法学に設定する．

憲法学に関する意味空間の形成においては，一例として憲法学の基本書である『憲法 第四版』[3]の章分けを参考とする．目次の構造に着目し，以下のようなグループ分けを行う．

- ◆第 1 部（1~4 章）⇒総論【大分野 X】（G1~G4）
- ◆第 2 部（5~13 章）⇒人権【大分野 Y】（G5~G13）
- ◆第 3 部（14~18 章）⇒統治【大分野 Z】（G14~G18）

「総論」「人権」「統治」といった憲法学の各サブ分野（大分野 X,Y,Z）の内容は，さらに，それぞれ 4 項目，9 項目，5 項目の小分野（G1~G18）に分類することができる．本方式では，これらの体系的知識の構造に着目し，それぞれを意味空間 Space X（4 次元），意味空間 Space Y（9 次元），意味空間 Space Z（5 次元）として構成する．

通常，文書データの特徴量を表すメタデータを用いた意味空間を作成するとき，着目する単語数とベクトル軸の数は等しくなり，数百に及ぶことも稀ではない．（[4]）

本論文では意味空間の形成のために，目次に，体系的・階層的構造を持つような知識源を用いて，その索引の参照ページから専門用語が大分野・中分野・小分野等のどの分野に関するものであるか，というデータを「単語分類知識ベース」として作成する．本稿の実験では，大分野と小分野の 2 段階の分類を採用している．

本方式で用いる単語分類知識ベースによって，意味的に関連する単語の「単語重要度」を集約し，ベクトル軸を整理する．本論文の実現システムにおいては，単語を 18 の小分野に分類しており，各文書に対し「18 次元のベクトルをもつメタデータ」を生成する．このメタデータを利用することによって，キーワードを用いた検索を行う従来の方式では困難な，意味的な関連性を考慮した検索のほか，特定分野に着目した検索や，全体的な意味を重視した検索に対応することができる．

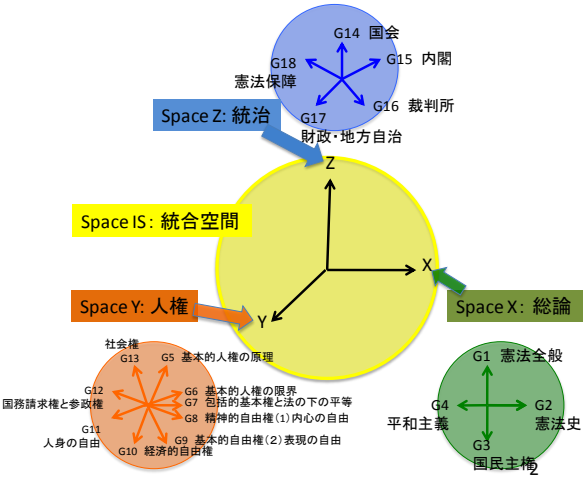

図 1　本複数意味空間連結機能を有する類似文書検索システムの基本アイデア

2.2. 目的に応じた意味空間の選択統合，および，文書のマッピング

本方式の特徴は，意味的関係性の計量による検索を実現するために，複数の意味空間を用意し，適用する空間を用途に応じて切り替え，複数の空間を統合する手段を提供することにある．

本方式では，文書群の各文書に含まれる専門用語に与えられた単語重要度(本論文では TF-IDF を用いる)と，専門用語の属する分野を定義したメタデータ（単語分類知識ベース）を利用し，分析対象となる文書を意味空間上にマッピングする。

本方式により，利用者の意図に応じた意味空間の切り替えが可能となる．これにより，「特定分野に着目した検索」と「全体的な意味を重視した統合的な検索」を選択することができる．

特定分野に着目した検索を行うことで，利用者の意図に沿った効率的な検索が可能になるほか，全体的な意味を重視した統合的な検索を行うことで，新たな知識の発見につながる検索を行うことができる．

2.2.1. 特定分野に絞った検索

特定の分野（例：「人権」）に絞った検索を行うときは，その分野に含まれる小テーマ（「人権」分野であれば後述の G5~G13）に関する単語の各文書における単語重要度を用いてマッピングし，他の分野に関しては考慮しない検索を行う．

2.2.2. 全体的な意味を重視した検索

特定の分野を指定せず，全体的に類似文書の検索を行うときは，「総論（X）」「人権（Y）」「統治（Z）」それぞれの大分野に含まれる小分野（X: G1~G4, Y: G5~G13, Z: G14~G18）のベクトルの長さを求め，後述

の正規化処理を行った結果として得られた値を用いて，X，Y，Z の各空間の要素を反映させた統合空間（IS 空間）にマッピングしなおし，相関量を計算する．

2.3. 相関量の計量

相関量の計量には内積（inner product）を用いる．計量の際は，各文書の情報量（文字数）の差異や出現する専門用語についての分野別の偏りに対応するために「3. 実現方式」の Step4, Step5 で説明する正規化処理を行う．

2.4. メタデータベースの定義

本方式では，文書間の相関量計量のために，以下のデータベースを定義する．

T：単語重要度データベース（Term Importance）
C：単語分類知識ベース（Classification）
S：単語重要度集計値データベース（Sum of Term Importance）
M：文書メタデータベース（Meta Database）

単語重要度データベース（*T*）は，各文書内における各単語の単語重要度（本論文では TF-IDF 値を用いる）を格納したデータベースである．

単語分類知識ベース（*C*）は，単語をどの分野に属するものかを定義する知識ベースである．今回対象とする法律学分野においては以下の 18 の小分野を用いて，各単語にたいして 18 次元のベクトルを与える．（小分野は以後 G1, G2, …, G18 のように表される．）また，それぞれの次元は 3 つの大分野 X, Y, Z に属する．大分野は，全体的な意味を重視した統合的な検索の際に用いられる．

本システムの分析対象となるデータは，法律学の一分野である「憲法」に関する文書群である．

専門用語がどの分野に属するのかは，憲法学において権威があるとされている基本書[3]の索引語リストとその参照ページのデータを利用した．図 2 に示す．

大分野	小分野	内容
(X) 総論	G1	憲法全般
	G2	憲法史
	G3	国民主権
	G4	平和主義
(Y) 基本的人権	G5	基本的人権の原理
	G6	基本的人権の限界
	G7	包括的基本権と法の下の平等
	G8	精神的自由権(1)内心の自由
	G9	精神的自由権(2)表現の自由
	G10	経済的自由権
	G11	人身の自由
	G12	国務請求権と参政権
	G13	社会権
(Z) 統治	G14	国会
	G15	内閣
	G16	裁判所
	G17	財政／地方自治
	G18	憲法の保障

図 2　検索・分析対象の憲法学関連文書項目

単語重要度集計値データベース（*S*）は，各文書において *C* を用いて各単語の単語重要度を分野別（G1～G18）に集約した結果を格納したデータベースである．

文書メタデータベース（*M*）は，各文書における小分野 G1～G4，G5～G13，G14～G18 の値をそれぞれ投影した X, Y, Z 軸を要素とする統合空間上にマッピングするためのメタデータを格納したデータベースである．

2.5. 基本機能

クエリとして文書データベース中の特定文書が指定されると，それに類似した文書の検索を実行する．

1）特定分野に着目した文書検索
2）全体的な意味が近い文書の検索

3. 実現方式

本方式により実現した法学分野文書対象意味検索システムの構成を図 3 に示す．

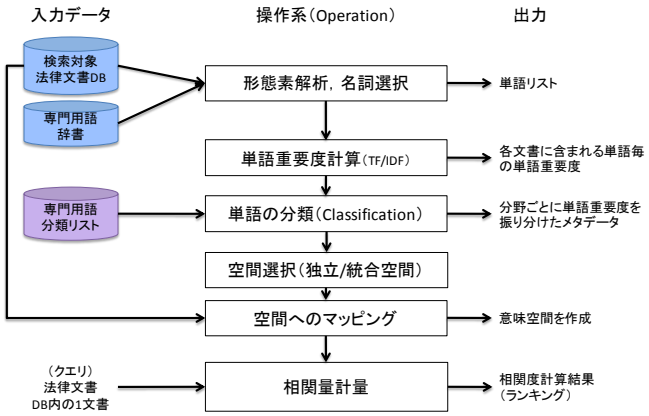

図 3　複数意味空間連結機能を有する法学分野文書対象意味検索システムの構成

本システムでは，以下の順に処理が実行される．

Step1　形態素解析

形態素解析には，オープンソースソフトウェアである MeCab[5]を使用する．形態素解析の際に，専門用語が分割されてしまうのを防ぐために，それらを MeCab のユーザ定義辞書に登録し，意図した形態素解析が行われるように設定した．

Step2　単語重要度（TF-IDF 値）の計算

本論文では，単語重要度の一例として TF-IDF 値を利用する．他の指標を用いることも可能である．TF（Term Frequency：特定文書における特定単語の頻出

度）と IDF（Inverse Document Frequency：文書群の中で特定単語を含む文書数）を考慮することで，その単語が各文書を特徴付けている度合いを数値化して求める．単語重要度データベース（*T*）を出力する．

Step3　単語重要度の分野別集計

専門用語を分野別に集計し，各文書に対する G1～G18 についての単語重要度の集計値を求める．その際には単語分類知識ベース（*C*）を用いる．具体的な処理方法に関しては以下の図を用いて説明する．

ただし，「word-」のあとにアルファベットの大文字が付与されているものは，単語分類知識ベースに含まれている単語である．これらの単語は図 4，図 5 において青色で強調されている．

また，アルファベットの小文字が付与されているものは，単語分類知識ベースに含まれていない単語で，それらの単語の単語重要度は図 6 に示すメタデータ生成には用いないものとする．

term	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16	G17	G18
word-A	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
word-B	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
word-C	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0
word-D	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
word-E	0	0	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0

図 4　単語分類知識ベース（専門用語の分類リスト）

file	term	tfidf
doc1	word-A	120
doc1	word-m	60
doc1	word-B	15
doc1	word-n	78
doc1	word-C	73
doc1	word-o	47
doc1	word-p	38
doc2	word-q	100
doc2	word-r	8
doc2	word-s	33
doc2	word-D	45
doc2	word-E	28
doc2	word-t	31
doc2	word-u	18

図 5　単語重要度リスト

file	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16	G17	G18
doc1	0	0	0	0	0	0	120	73	120+15+73	0	0	0	73	0	0	0	0	0
doc2	0	0	28	0	28	28	28	0	0	0	0	0	45+28	0	0	0	0	0

図 6　各文書について生成される 18 次元のベクトルをもつメタデータ

(1) 単語分類知識ベースに含まれる単語の重要度

図 5 においては，文書 doc1，doc2 に含まれる単語のうち，図 4 の単語分類知識ベースに含まれる単語が青色で強調されている．まず，一例として文書 doc1 に含まれる単語である word-A について考える．図 5 より，word-A は単語重要度 120 であることが分かる．

(2) 単語と各分野との関係性

図 4 では，各単語について，関係性がある分野との組み合わせにおいて「1」が付与されている．これより，word-A は G7 と G9 に関係のある単語であることがわかる．

(3) 対応する分野に，単語重要度を加算

図 6 の各文書のメタデータ生成において，行 doc1 の列 G7,G9 にそれぞれ 120 を加算する．

以上の(1)～(3)の処理を各文書の単語それぞれにおいて繰り返し実行し，メタデータの生成を行う．

Step4　2 ノルム正規化

各文書ベクトル（G1～G18 を要素とするベクトル）の長さを 1 とする，2 ノルム正規化を行い，各文書の情報量の違いによる検索結果への影響をなくし，各文書を対等に扱うように設定する．

Step5　無限大ノルム正規化

無限大ノルム正規化は，全体的な意味を重視した検索を行う際に行う．

大分野 X,Y,Z について，それぞれに属する小分野 G1~G4，G5~G13，G14~G18 の単語重要度によって表されるノルムの大きさを求める．この値を用いて各文書を X,Y,Z を軸とした統合空間にマッピングする．

X,Y,Z を軸とした統合空間において，X,Y,Z の値は各文書のそれぞれの分野（総論，人権，統治）における寄与度を表している．

各文書を対等なものとして扱うために，各分野 X,Y,Z について，文書群中の全文書のうち最大の寄与度を表す値を「1」として正規化を行う．これにより，X,Y,Z 各分野を対等に扱い，X,Y,Z の間の情報量の違いが検索結果に影響を与えないようにする．

文書群の内容が特定分野の内容に偏っている場合，全体的に値が小さい分野が，相関量計算の結果に現れにくくなる場合がある．

例えば，検索対象となる文書群において Y についての記述が多い場合を考えてみる．その際，内積による相関量計算を行うと比較文書間の X と Z における積は Y についての積に比べ小さくなり，X と Z についての相関度への影響が小さくなってしまう．これは，X と Z の重みを小さく与えていて，Y に重みを大きく与えていることと同じことである．Step5 の処理によって，X と Z の相関度への影響度の小ささを是正することができる．

無限大ノルム正規化の例を図 7，図 8 に示す．

	X	Y	Z
doc1	0.2	5	0.2
doc2	0.5	8	0.1
doc3	0	3	0.2
最大	0.5	8	0.2

図7　X,Y,Zのノルムの長さおよびX,Y,Zそれぞれの最大値

	X	Y	Z
doc1	0.4	5/8 (≒0.6)	1
doc2	1	1	0.5
doc3	0	3/8 (≒0.3)	1

図8　図7の無限大ノルム正規化処理の結果

例えば，X,Y,Z のノルムの長さを求めた結果，図 7

のようになった場合，X の最大値は 0.5，Y の最大値は 8，Z の最大値は 0.2 である．X,Y,Z の最大値をそれぞれ 1 とするために，全ての文書の X,Y,Z をそれぞれ 0.5，8，0,2 で割る処理を行う．その結果である図 8 に示された値を用いて、全体的な意味を考慮した統合的な相関度計算を行う．

Step6　相関量の計算

特定分野に着目した検索を行う際は Step4 までを行い，全体的に近い意味をもつ文書を検索する際は Step5 までを行い，計量を行う．

本論文では，各文書は 18 次元のベクトルで表される．ユーザによって選択された分野の値について内積を利用して比較している．

特定分野に絞った検索を行う際は，指定した分野に含まれる小分野の単語重要度のみを用い，他の分野は考慮しないこととする．例えば，X 分野のみに着目して検索を行う際には，18 次元のベクトルのうち，G1~G4 の 4 つの分野の値を利用して相関量計算を行う．

全体的な意味を重視した検索を行う際は，G1~G4, G5~G13,G14~G18 それぞれからなるノルムの大きさをそれぞれ，総論（X），人権（Y），統治（Z）の値とし，X,Y,Z それぞれの分野において Step5 に示した無限大ノルム正規化を行い，X,Y,Z を軸とした統合空間をつくり，相関量を計算する．

4. 実験

複数の意味空間の連結により，利用者の目的に応じた類似文書の検索を行う意味検索システムについて，実験により実現可能性を評価する．

4.1. 実験環境

実験に用いた文書群は次のとおりである．

1. SFC 霞会[*1] 憲法勉強会レジュメ[6]
   [*1] 慶應義塾大学の公認サークル
2. 憲法学の解説書[7]

本実験では，[3]の索引語（883 語）および参照ページを利用した分類（Classification）リストを使用する．この索引語には専門用語と一般語の両方が含まれる．次にその例をあげる．ただし，索引中ではその区別はなされていないため，一般的な例を挙げるものとし，本実験では索引語すべてが利用されるため，その区別は実験の結果には影響は及ぼさない．

- **専門用語の例**：「旭川学テ事件」等の固有名詞，「立法裁量」，「議院自律権」，「明白かつ現在の危険の基準」等の複合語，その他．
- **一般語の例**：「軍隊」，「基地」，「警察」，「元首」，「死刑」，「戦争」，「選挙」，「保障」など．

実験は次の手順により実施する．

実験 1　文書「jinken-05-seishin.txt」に類似する文書を，**各分野のみに着目**して検索を行う．

実験 1-X　**大分野 X（総論）**（小分野 G1~G4）
実験 1-Y　**大分野 Y（人権）**（小分野 G5~G13）
実験 1-Z　**大分野 Z（統治）**（小分野 G14~G18）

実験 2　文書「jinken-05-seishin.txt」に類似する文書を，**大分野 X,Y,Z すべてを考慮した全体的な意味を重視した統合的な検索**を行う．（G1～G18 全てを用いる）

実験 2-A　類似文書を，**大分野 X,Y,Z の各分野に着目してそれぞれ 1 回ずつ計 3 回の検索を行い，その 3 回の結果の総合的評価（相関量に応じて並べ替え）により**全体的な意味を重視した統合的な検索を行う．

実験 2-B　類似文書を**本論文の提案方式によって正規化し**，全体的な意味を重視した統合的な検索を行う．（G1~G18）

4.2. 実験結果

実験結果のうち，上位 10 位までに該当する文書を図 9 から図 13 に示す．

実験 1-X

順位	文書名	相関度
1	jinken-10-shakaikei.txt	0.011369775
2	b07-houkatsukihonken.txt	0.010302152
3	b08-hounomotonobyodo.txt	0.009792159
4	jinken-05-seishin.txt	0.009517622
5	jinken-09-jinshin.txt	0.009490119
6	jinken-04-houkatsubyodo.txt	0.007183232
7	jinken-07-keizai.txt	0.006854621
8	b09-hyogenjiyuu.txt	0.006600507
9	b16-sanseiken.txt	0.005001321
10	jinken-06-hyogen.txt	0.004209992

図 9　指定文書（jinken-05-seishin.txt）に類似している文書を大分野 X（総論）のみに着目して検索した結果

実験 1-Y

順位	文書名	相関度
1	jinken-05-seishin.txt	0.985885513
2	b10-shisouryoshin.txt	0.951648164
3	b11-shinkyojiyuu.txt	0.950772053
4	b13-kyoikutokenpou.txt	0.343282525
5	jinken-06-hyogen.txt	0.144762435
6	b05-jinkengenri.txt	0.127857119
7	jinken-03-kihonteki.txt	0.125495157
8	b06-jinkenseiyaku.txt	0.118380551
9	b09-hyogenjiyuu.txt	0.108007759
10	jinken-10-shakaikei.txt	0.107492080

図 10　指定文書（jinken-05-seishin.txt）に類似している文書を大分野 Y（人権）のみに着目して検索した結果.（人権に関する文書（あずき色で強調されている箇所）以外にも，思想や表現の自由に関する文書（黄色で強調されている箇所）が関連する文書として検索できていることが分かる.）

実験 1-Z

順位	文書名	相関度
1	jinken-00-intro.txt	0.041748781
2	b06-jinkenseiyaku.txt	0.024644644
3	jinken-02-jinkengenri.txt	0.024019980
4	b08-hounomotonobyodo.txt	0.022696671
5	b05-jinkengenri.txt	0.018679397
6	jinken-08-zaisankokumusansei.txt	0.017690554
7	b16-sanseiken.txt	0.017430387
8	jinken-03-kihonteki.txt	0.017202230
9	b02-kenpoushi.txt	0.016132139
10	jinken-01-souron.txt	0.015546007

図 11　指定文書（jinken-05-seishin.txt）に類似している文書を大分野 Z（統治）のみに着目して検索した結果.（実験 1-Y の結果中の人権に関する文書（あずき色で強調されている箇所）以外にも，統治に関する文書が検索できていることが分かる.）

実験 2-A

実験 1-X, 1-Y, 1-Z の結果の総合評価

順位	文書名	相関度
1	jinken-05-seishin.txt	1.000000000
2	b10-shisouryoshin.txt	0.978742210
3	b11-shinkyojiyuu.txt	0.976160744
4	b13-kyoikutokenpou.txt	0.401767365
5	b06-jinkenseiyaku.txt	0.187986439
6	b05-jinkengenri.txt	0.186156602
7	jinken-03-kihonteki.txt	0.179156886
8	jinken-06-hyogen.txt	0.152038213
9	jinken-02-jinkengenri.txt	0.146654887
10	jinken-10-shakaikei.txt	0.126014375

図 12　指定文書に類似している文書を，大分野 X,Y,Z の各分野に着目してそれぞれ 1 回ずつ計 3 回の検索を行い，その 3 回の結果の総合的評価（相関度に応じて並べ替え）により全体的な意味を重視した統合的な検索を行った結果.

実験 2-B

順位	文書名	相関度
1	jinken-10-shakaikei.txt	1.018130611
2	b09-hyogenjiyuu.txt	1.017580533
3	jinken-07-keizai.txt	1.017574487
4	b07-houkatsukihonken.txt	1.017558303
5	jinken-05-seishin.txt	1.017475919
6	jinken-06-hyogen.txt	1.017260394
7	jinken-04-houkatsubyodo.txt	1.016738213
8	b14-jinshinnojiyuu.txt	1.014119245
9	jinken-09-jinshin.txt	1.010634105
10	b10-shisouryoshin.txt	1.006348001

図 13　指定文書に類似している文書を本論文の提案方式によって正規化し，全体的な意味を重視した統合的な検索を行った結果.（実験 2-A の結果と比較すると，実験 2-A の 2 位と 10 位の文書が，実験 2-B ではそれぞれ 10 位と 1 位に出現し,順位が逆転していることが分かる.）

4.3. 考察

指定した問い合わせについて，ランキングした上位 10 文書について着目していく.

4.3.1. 特定分野に着目した検索に関する考察

実験 1-Y において，問い合わせとして指定した文書である「jinken-05」（精神的自由権に関する文書）に関連度の高い，「b10（思想・良心の自由に関する文書）」，「b11（信仰の自由に関する文書）」，「jinken-06（表現の自由に関する文書）」，「b09（表現の自由に関する文書）」といった文書が全 27 文書中のうち，9 番目までに出現しており，特定分野に限定した本検索に有効性があることが読み取れる.

実験 1-Z について，実験 1-Y と比較すると実験 1-Z の第 2 位，5 位，8 位のみに実験 1-Y と同じ文書がランキングされた．これより，異なった視点による検索が行われていることが読み取れる.

なお，実験 1-X，実験 1-Y，実験 1-Z において，問い合わせとして設定した文書自身との相関度が 1 となっていないのは，18 次元で表される文書ベクトルの長さが 1 となるような正規化を行い,その 18 次元のうち一部（Y 分野ならば G1~G18 のうち G5~G13 のみ）を選択して相関量を計算しているためである.

4.3.2. 全体的な意味を重視した検索に関する考察

実験 2 においては，実験 2-A と実験 2-B の実験を行った.

実験 2-A では，X 分野，Y 分野，Z 分野それぞれに着目した 3 回の検索を行い，それぞれ計量がなされた相関量の合計の値について降順にソートしたものを総合的評価として用いる.

実験 2-Bでは,本論文の提案方式によって正規化し,全体的な意味を重視した検索を行った.

実験 2-Aと実験 2-Bを比較すると,「jinken-10」は実験 2-Aでは 10 位であるものの,実験 2-Bにおいては 1 位にランキングされており,また対照的に「b10」は実験 2-Aでは 2 位であるものの,実験 2-Bでは 10 位にランキングされている.

これにより,本論文の提案方式による正規化によって,異なった検索結果を得ることができた.

実験 2 において,全体的な意味を捉えた検索を行ったが,検索結果を検討すると,本論文の提案方式による正規化(無限大ノルム正規化)を行った実験 2-Aよりも,各文書につき,大分野 X,Y,Z の各分野のみに着目して検索を行った実験 2-Bのほうが正確な結果が得られている.この原因については,検索対象となるサンプル数が少なかったなどの原因が考えられ,サンプル数を増やし,再検証を行う必要がある.

なお,実験 2-Bにおいて問い合わせとして設定した文書自身との相関度が 1 となっていないのは,実現方式の Step5 による無限大ノルム正規化によって,X,Y,Z それぞれの最大値を 1 とする正規化を行い,各文書の X,Y,Z に重みづけを行っているためである.

また,実験 1 および実験 2 で行っている実現方式の Step4 にある 2 ノルム正規化については,検索結果に文書の情報量に差がある「jinken-」から始まる文書(レジュメ)と「b-」から始まる文書(書籍の一部)が混在しており,偏りがないことから,有効性があると認められる.

4.3.3. 結果の分析に関する考察

本実験の結果の分析は,法律学を学んでいる著者によるものであるが,検索の有効性をはかる客観的な指標の検討が必要である.指標としては,法律学を学んでいる者や法律に詳しくない者を含む被験者を対象として,検索結果の妥当性を判断してもらう方式が考えられる.

5. おわりに

本稿では,複数の意味空間の連結により,利用者の目的に応じた類似文書の検索を行う意味検索システムを示した.

本システムは,利用者の目的に応じた類似文書の検索を行うことができる.

本稿では,文書群の各文書に対して単語分類知識ベースを適用することにより,類似文書の検索実験を示した.

実験により,本システムによって,特定分野に着目した文書検索,全体的な意味を捉えた文書検索を行うために,複数の意味空間を検索目的に応じて使用することが可能であることを示した.

本システムは,他の分野の文書群にも適用可能であり,様々な文書に関して効率的な検索を行うことができる可能性がある.

今後は,オントロジーなどを用いた他の手法との比較や,本論文の提案方式について,それぞれの分野ごとの意味の近接度に応じた重み付けや,全体的な意味検索の結果に対する客観的指標の可能性,およびそれらの精度向上の手段について検討したい.

参考文献等

[1] LEX/DB インターネット　TKC 法律情報データベース(http://www.tkclex.ne.jp/).

[2] 裁判所判例検索システム(http://www.courts.go.jp/).

[3] 芦辺信喜著,高橋和之補訂『憲法 第 4 版』,岩波書店, 2007.

[4] 佐々木 史織,清木 康,薬師寺 泰蔵,“国際関係分野ドキュメント群を対象とした意味的連想検索のための空間生成方式”,日本データベース学会 Letters, Vol.2, No.1, pp.39-42, 2003.

[5] MeCab(京都大学情報学研究科−日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトによる. http://mecab.sourceforge.net/)

[6] SFC 霞会憲法勉強会レジュメ.(SFC 霞会は慶應義塾大学の公認サークル)

[7] 木村哲也『初学者のための憲法学』,北樹出版, 2008.3, pp.14-221.