修 士 論 文

Fタームによる特許分類のためのカーネル設計

北陸先端科学技術大学院大学
知識科学研究科知識科学専攻

三浦 祥治

2011年3月

修 士 論 文

F タームによる特許分類のためのカーネル設計

指導教員　Ho Tu Bao 教授

北陸先端科学技術大学院大学
知識科学研究科知識科学専攻

0850031 三浦 祥治

指導教員:　Ho Tu Bao 教授 (主査)
橋本 敬 教授
藤波 努　准教授
由井薗 隆也　准教授

提出年月: 2011 年 2 月

目 次

第1章　序論

1.1　特許の社会的背景

近年の企業競争のグローバル化や知的財産権の有効活用に伴って，企業の競争力の強化を図る戦略の一つとして，特許を取得する活動が激しくなっている．知的財産権に対する企業の活動の背景として， 2002 年 7 月に日本政府が発表した「知的財産戦略大綱」がある．この「知的財産戦略大綱」では，知的財産立国を目指し，その戦略として，知的財産の「創造戦略」,「保護戦略」,「活用戦略」,「人的基盤の充実」の四つを掲げている [1, 2]．また,「保護戦略」の一つとして,「迅速かつ的確な特許審査・審判」が挙げられており，審査体制の整備や国際的協調を含む総合的な対策が必要であると示されている [2]．2001 年に特許審査請求期間が 7 年から 3 年に短縮されたことで特許審査請求件数が年々増加し，特許行政年次報告書の 2009 年版 [3] によれば，図 1.1 からわかるように，日本国への特許出願件数は，年間 40 万件前後となっている．

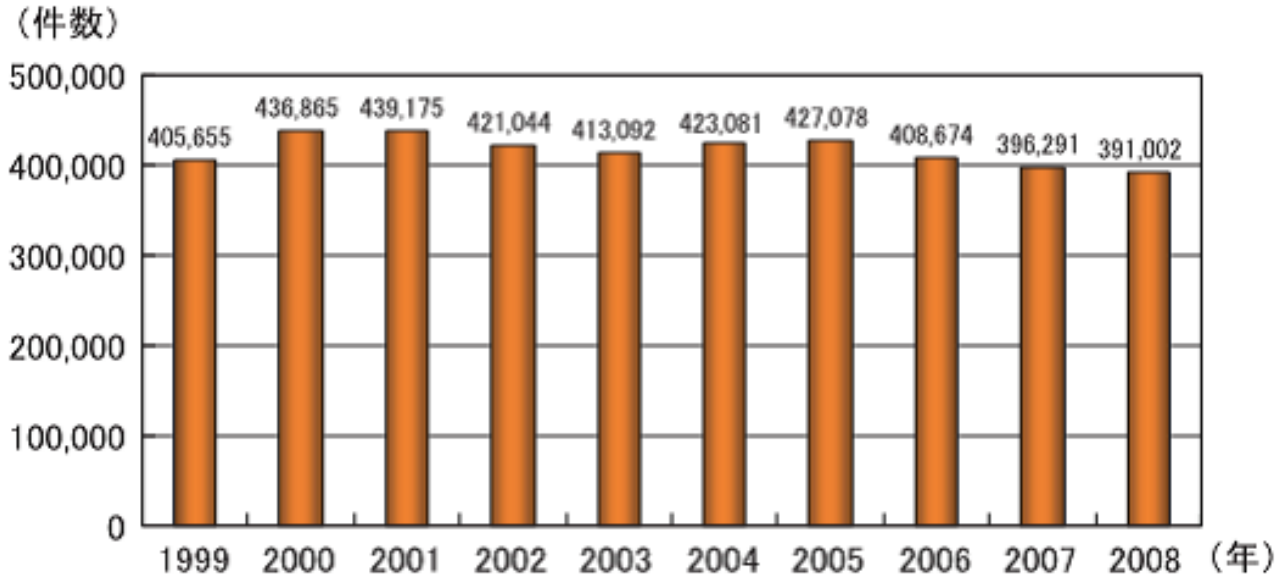

図 1.1: 特許出願数の推移

また，図 1.2 に示すように，日本国における特許審査官一人あたりの特許審査の処理件数は他地域に比べ非常に多い．この図 1.2 からわかるように，日本国における特許審査官

は一人当たりに対する処理件数が非常に多く，その結果，特許審査官には負担がかかり，出願された特許の処理の件数が減少しない状態が続いた．そこで，特許庁は審査処理能力を向上させるため，2004 年から 2008 年までの 5 年間で約 500 名の審査官の増員が行われてきた．その毎年の審査官の増員の結果，徐々に審査官による特許の審査件数が増えている．しかしながら，北米，欧州に比べ，日本国における一人の特許審査官あたりの審査処理件数が多い状況が変わらず [3]，審査官にかかる特許審査の処理の負担が減らないため，図 1.3 に示すように，審査請求から審査開始までの審査待ち期間が 28 ヶ月と長くなっており，それとともに審査待ちの特許数が減少しないことから，特許庁における特許審査業務を圧迫している [4].

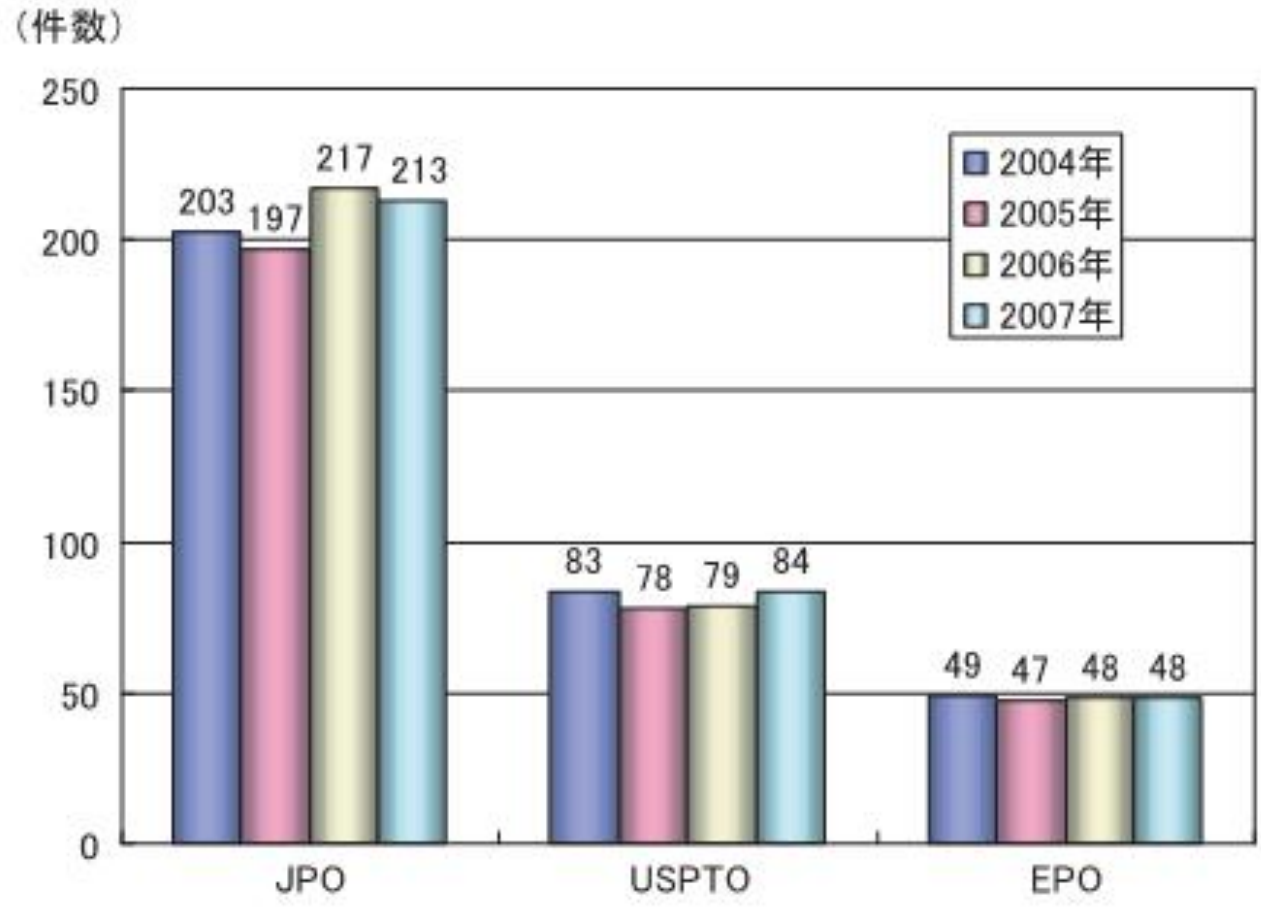

図 1.2: 審査官あたりの特許数のグラフ

特許は網羅する技術分野が非常に広く，審査には技術分野の専門知識及び特許文書検索ノウハウを要するため，審査官の増員も容易ではない [2]．さらに，知的財産権の戦略的活用を模索している企業内においても，製品やサービスに係る特許の戦略的な取得・活用

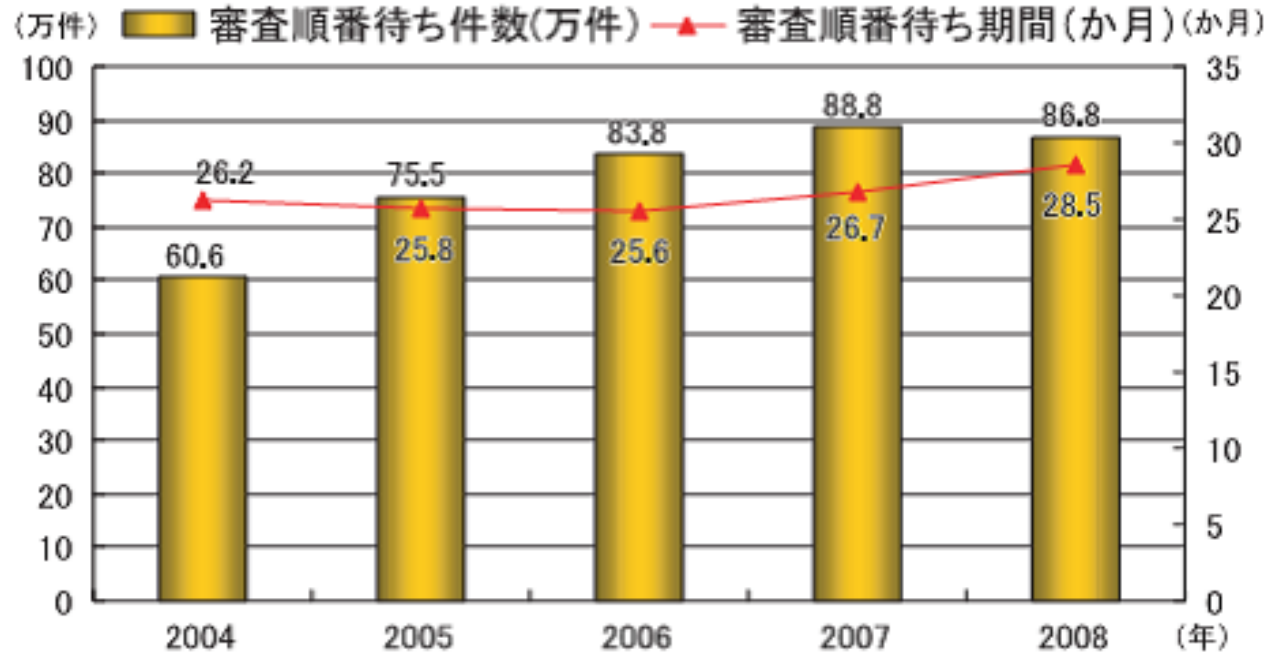

図 1.3: 審査順番待ち件数と審査順番待ち期間に関するグラフ

と，特許管理の低コスト化が大きな課題となっており，特許の戦略的活用の促進を支援する計算機システムへの期待が高まっている [2].

1.2 計算機システムの利用

前節で解説した社会的背景から，計算機システムを利用した特許の検索，調査，分析を行うシステムの研究が行われるようになった．しかし，特許文書検索は当初は特許の出願が紙ベースで行われていたため，検索できる特許は限りがあり，アクセスできる人も限られていた [2]．しかし，テキスト文書のデジタル化が行われるようになり，特許の出願においてもデジタル化が行われるようになった．この結果，特許文書をデジタルデータとして蓄積できるようになり，1999 年に特許電子図書館である IPDL(Industrial Property Digital Library) のサービスが開始され，一般の人もインターネット経由で簡単にアクセスすることが可能になり特許文書検索の効果的な方法などが研究されるようになった．

1.3 特許検索

計算機システムの利用が可能になり，一般の人も利用可能になった結果，特許を検索する出願特許の内容を理解するシステムがあげられる．このようなシステムで特許を検索す

る際，多くのユーザーは文書検索を行う．そして，この検索で最も主流となっている検索方法はキーワードによる論理式検索である [2]．この検索方法は AND/OR/NOT などの論理演算子を用いてキーワードの論理式を表現し，検索に用いる．このように計算機システムを使用した特許の検索や調査，分析に関する研究が行われている．これらの研究において，NTCIR(NII-NACSIS Test Collection for IR Systems) という情報検索やテキスト要約・情報抽出などのテキスト処理技術の研究のさらなる発展を図るワークショップ型共同研究のプロジェクトにおいて，特許分類に関する研究が行われている．

1.4 特許分類の先行研究

NTCIR の研究の中に「F ターム」に焦点を当てた分類タスク [5, 6] があり，その研究の多くで，機械学習で有名かつ有効なアルゴリズムである Suppoer Vectore Machine (SVM) を利用した研究が行われている．これらの研究において，特許の文書構造と文書の特徴に焦点をあて，それらの特徴に最適な処理を行ない，SVM を用いた F タームの自動付与の制度を向上させる研究を行っている [7, 8, 9]．

1.5 カーネル手法

SVM を用いた学習方法の中に，SVM の学習方法を改良するカーネル手法というものがある．この手法は与えられた各データのペア間の内積を計算し，行列を作る．その行列を用いて一般的な SVM による分類と同様に処理を行ない分類させる手法である．
カーネル手法の中に文書分類に適した手法が考えられている [10, 11, 12]．ベクトル空間モデルとして表現し，そのデータにカーネル手法を使う方法が一般的に行わている [10, 12]．この手法は文書分類において一般的な手法であり，かつ，有効な手法の一つである [13, 14]．

1.6 テキスト情報の分類

テキスト情報のデジタル化とインターネットの普及により，テキスト情報が急速に多くなり，もはや人手で分類されることは不可能なほどになっている [12]．この結果，世界的に文書の自動処理を行うことが人工知能や計算機科学において研究が行われるようになり，その研究分野でも，情報検索 (IR，information retrieval) に関しての研究が盛んに行われている [10, 11, 12]．テキスト分類の研究分野において，重要なポイントとなるのは，提案した方式の有効性を評価する環境の構築が問題となっていた [2, 15]．この問題に対して，米国では 1992 年に TREC(Text REtrieval Conference) や，2000 年に欧州で始まった CLEF(Cross Language Evaluation Forum) 1996 年に公開された日本語情報検索システム評価用テストコレクションの BMIR-J1 (BenchMark for Information Retrieval systems for

Japanese texts ver.1)，1998 年に国立情報学研究所などが主催で始まった情報検索システム評価用テストコレクション構築プロジェクトの NTCIR などの研究用の評価データセットを整えたことで，テキスト分類の研究が盛んになった [2].

1.7 本稿の目的

本稿は，毎年の大量の特許申請に対して審査官の早急な増員ができない中で，どのようにして審査待ちの期間を短くし，審査待ち件数を減少させるかという問題に対して，計算機を用いた解決を図る試みの一つとして特許を分類する研究を行われている．特許を分類するにあたり，特許が持っている特徴である分類システムに着目した．特許を様々な技術に分類する F タームというコードが特許には付与されている．このコードを利用することで審査官もしくは，特許の先行調査を行う企業の従業員などが特許の先行調査を行う際の特許の数を絞り込むことが容易になると考えられる．そのため，先行研究として F タームを利用した特許の自動分類の研究が NTCIR のデータ提供により盛んに行われている．NTCIR から提供されているデータの中で本稿が利用するデータは「NTCIR-6 特許検索文書データ」である日本国公開特許公報全文データ 1993 年から 2002 年までのデータを集めたものを利用している．先行研究では分類を行う際に SVM という学習器を使用し，特許や F タームの特徴を反映した SVM のアルゴリズムの改良を行うことで正確な F タームによる特許自動分類を目指している．また，先行研究において，K-Nearest Neighbor のアルゴリズムを使用しての研究も行われているが SVM を利用した研究と同じ評価を得ており，両者に差がみられていない [16]．特許の情報のほとんどが文章・テキストである点に注目することで，文章やテキストの分類を行う研究で有効な結果を導いているカーネル手法を特許自動分類にも適用されている．カーネル関数は最適なパラメーターを設定することでことで分類の精度の向上を図ることが可能であると考えられる．もともとカーネル手法は SVM の機能を向上させるひとつのアルゴリズムである．そこで従来の研究と同様にカーネル手法を適用し，カーネル手法の最適なパラメーターの組み合わせを特定のデータに限定し，そのパラメーターを発見し、カーネル手法を特許の自動分類に適応できる可能性を探り，また，カーネル手法による分類の精度の向上を目指す.

1.8 本稿の構成

本稿では，まず，2 章において，特許の特徴について，特許文書の構成と特許分類体系，F タームによる特許分類について解説を行い，また，本稿で用いるデータに付いても解説を行い，3 章では先行研究の方法論であるカーネル手法と Support Vector Machine (SVM) について解説を行い，カーネル手法と文書の分類との関係について解説し，4 章では 3 章で解説したカーネル手法を使用した F タームの分類の方法，評価方法を解説し，5 章では実際に実験に使うデータの解説，データをどのように加工するかという前処理と，実験環

境，実験方法，実験結果，そして実験結果に対する分析についてを解説し，6 章では本稿の目的に対して方法，実験の有効性について解説し，また，本稿における問題点，課題点などの考察を行いまとめる．

第2章　特許

本章では，本稿の内容の理解を助けるために，特許出願の構成と特許分類体系，そして，F タームによる特許分類について解説し，また，本稿で用いる特許のデータについても解説をする．

2.1　特許出願の構成

特許権を取得するためには、「特許願」及び権利を取りたい技術内容を詳しく記載した「明細書」,「特許請求の範囲」,「図面」,「要約書」を作成し、特許庁に出願する必要があります。特許を出願するには，決まった様式に則って出願をしなければならない．出願された特許は特許庁で審査される．特許出願後，出願日や分類などの属性データが専門家によって付加され，公開特許公報として一般に公開される．
特許を出願する際には，必要な要項があり，それらは半構造化されている．特許出願の必要要項は，発明の名称，請求項，技術分野，従来技術，発明が解決する課題，課題を解決する手段，発明の実施の形態，発明の効果，図面の簡単な説明，符号の説明，要約という構成になっている．

図 2.1 は実際の特許の構成の例である．そして，図 2.1 に関する各項目の説明 [2, 17] をする．

a:コード
　特許に付与されている IPC，FI，F タームのコードである。この特許においては，IPC が 2 つ，FI が 2 つ，F タームが 5 つ付けられている．

b:発明の名称
　発明の内容を簡潔，明瞭に表示する名称をつけ，発明の内容と関係のない字句を入れてはいけない．そのため，名称はその発明の対象物に関することを記載することが多く，例えば,「ロボットの二足歩行装置」や「電気自動車の充電制御方法」などである．

c:要約
　発明の目的・課題・手段について簡潔に記載している．

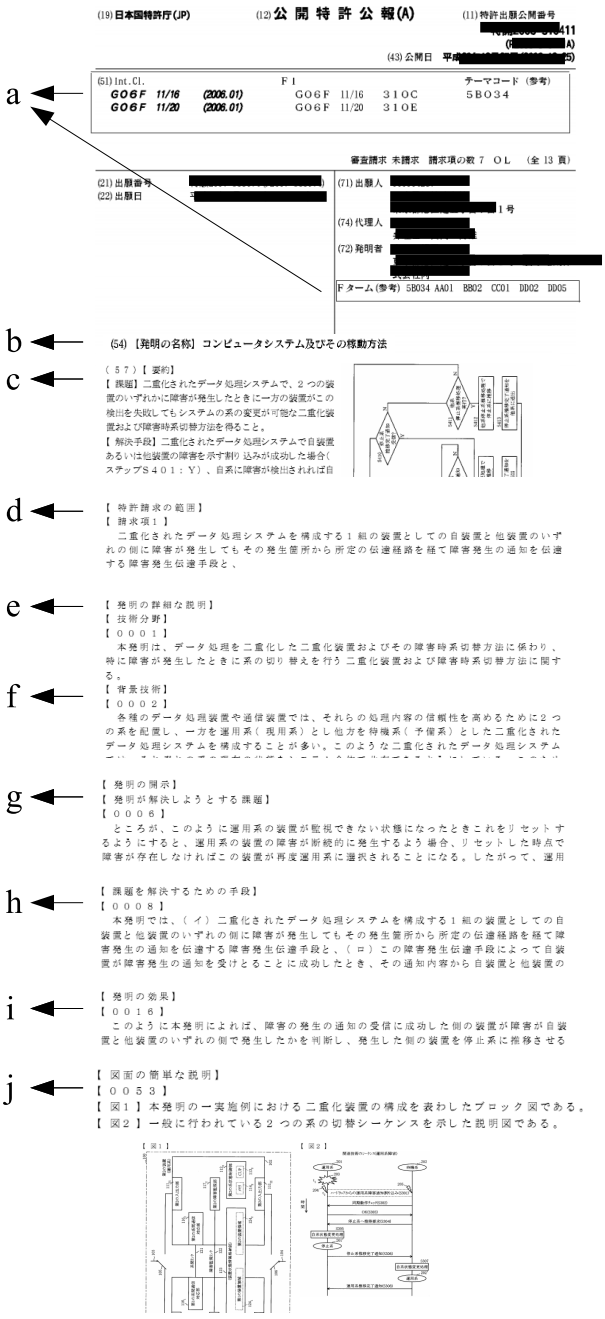
a
b
c
d
e
f
g
h
i
j
(19)日本国特許庁(JP)
(12)公開特許公報(A)
(11)特許出願公開番号
(43)公開日
(51)Int.Cl.
F I
テーマコード（参考）
G06F 11/16 (2006.01)
G06F 11/20 (2006.01)
G06F 11/16 310C
G06F 11/20 310E
5B034
審査請求 未請求 請求項の数 7 ＯＬ （全 13 頁）
(21)出願番号
(22)出願日
(71)出願人
(74)代理人
(72)発明者
Fターム(参考) 5B034 AA01 BB02 CC01 DD02 DD05
(54)【発明の名称】コンピュータシステム及びその稼動方法
（５７）【要約】
【課題】二重化されたデータ処理システムで、２つの装置のいずれかに障害が発生したときに一方の装置がこの検出を失敗してもシステムの系の変更が可能な二重化装置および障害時系切替方法を得ること。
【解決手段】二重化されたデータ処理システムで自装置あるいは他装置の障害を示す割り込みが成功した場合（ステップＳ４０１：Ｙ）、自系に障害が検出されれば自
【特許請求の範囲】
【請求項１】
二重化されたデータ処理システムを構成する１組の装置としての自装置と他装置のいずれの側に障害が発生してもその発生箇所から所定の伝達経路を経て障害発生の通知を伝達する障害発生伝達手段と、
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、データ処理を二重化した二重化装置およびその障害時系切替方法に係わり、特に障害が発生したときに系の切り替えを行う二重化装置および障害時系切替方法に関する。
【背景技術】
【０００２】
各種のデータ処理装置や通信装置では、それらの処理内容の信頼性を高めるために２つの系を配置し、一方を運用系（現用系）とし他方を待機系（予備系）とした二重化されたデータ処理システムを構成することが多い。このような二重化されたデータ処理システム
【発明の開示】
【発明が解決しようとする課題】
【０００６】
ところが、このように運用系の装置が監視できない状態になったときこれをリセットするようにすると、運用系の装置の障害が断続的に発生するよう場合、リセットした時点で障害が存在しなければこの装置が再度運用系に選択されることになる。したがって、運用
【課題を解決するための手段】
【０００８】
本発明では、（イ）二重化されたデータ処理システムを構成する１組の装置としての自装置と他装置のいずれの側に障害が発生してもその発生箇所から所定の伝達経路を経て障害発生の通知を伝達する障害発生伝達手段と、（ロ）この障害発生伝達手段によって自装置が障害発生の通知を受けとることに成功したとき、その通知内容から自装置と他装置の
【発明の効果】
【００１６】
このように本発明によれば、障害の発生の通知の受信に成功した側の装置が障害が自装置と他装置のいずれの側で発生したかを判断し、発生した側の装置を停止系に推移させる
【図面の簡単な説明】
【００５３】
【図１】本発明の一実施例における二重化装置の構成を表わしたブロック図である。
【図２】一般に行われている２つの系の切替シーケンスを示した説明図である。
【図１】
【図２】

図 2.1: 特許の構成の例

d:請求項

特許を受けようとしている発明を特定するために必要とする内容について記載している．一つの請求項には一つの発明内容を記載している．請求項において，それより前に記載された他の請求項を引用して記載した，従属請求項と，他の請求項を引用せずに記載した独立請求項とに分かれている．

e:技術分野

特許を受けようとしている発明の技術分野について明確にし，その内容を簡潔に記載している．

f:背景技術

特許を受けようとする発明に関連する従来の技術，またはすでに開発されている技術の内容についても記載している．

g:発明が解決しようする課題

特許を受けようとする発明が課題にしている従来技術の問題点などを記載している．

h:課題を解決するための手段

請求項に記載されている内容が課題の解決手段となるので，請求項の構成を記載している．

i:発明の効果

特許を受けようとする発明が，従来の技術と比べ優れている点，発明の有利な効果などについての解説を記載する．この項は発明の進歩性を判断する材料となるため重要な項である．

j:図面の簡単な説明

説明に用いられている各図面の簡潔に記載している．図面タイトルと同様である

2.2 特許に関する特徴

特許に関する特徴として3つが挙げられている [2]．1つ目は，多くの発明者が特許文書を執筆していることで，文章記述スタイルや文章長，使用語彙 (異表記，同義語も含む) が様々であること．2つ目は，特許は多岐にわたる技術分野を網羅していることで，分野ごとに執筆スタイルが異なる．例えば，化学分野では化学式が頻繁に使われ，機械分野では図面の使用が頻繁にあるなどである．3つ目が発明者と出願人の関係についてである．発明者はその発明を考案するのに貢献した人々を指し，出願人はその発明を特許として権利化することを申請する人または組織である．ほとんどは，発明に最も貢献した発明者が執筆するが，特許事務所や企業内の知財部の専門家によって執筆されることもある．

2.3　日本の特許分類体系

現在，特許はその技術内容，特徴によって，それぞれ分類がつけられている．世界共通で使用されている国際特許分類 (Internationnal Patent Classification，IPC) というのがある．世界共通の特許分類ではあるが，各国，各地域で独自の特許分類の開発を行っている．例えば，アメリカ独自の米国特許分類 (U.S. Patent Classification，USC) や欧州独自の欧州国際分類 (European Patent Classification，EPC もしくは ECLA）や日本独自の File Index (FI)，F タームなどがある．本稿では IPC，FI，F タームのみの解説をする．

2.3.1　国際特許分類 (IPC)

IPC は, 特許文献のおく最適に統一した分類を得るための手段であり，特許出願中の技術開示について，新規性，進歩性または非自明性を評価するために，知識財産庁や他の利用者が特許文献を検索するための有効なサーチツールの確立を目的としている [?, ?]．さらに IPC は，次の４つの重要な目的を持っている．1 つ目は，特許文献に含まれている技術及び権利情報へ容易にアクセスするための特許文献の秩序だった整理のための道具となること．2 つ目は，特許情報のすべての利用者に情報を選択的に普及させるための基礎となること．3 つ目は，ある技術分野における技術の状況を調査するための基礎となること．4 つ目は，種々の分野における技術の発展をも評価できる工業所有権についての統計を作成するための基礎となること [24]，という目的をもっている．
IPC は生物分類のように下の階層に行くほど詳しく分類がされるようにできており，上位の階層から順に，セクション，クラス，サブクラス，メイングループ，サブグループから構成されている．これらは階層構造になっており，最も高い階層はセクション，次に第 2 階層のクラス，その次に第 3 階層がサブクラス，その次の第 4 階層がメイングループ，その次の第 5 階層がサブグループという 5 階層になっている．以下は各階層の説明である

セクション
: 特許の分野に適当であると認められる知識体系を 8 つにわけている．それらをセクションと呼び，そのセクションは大文字の A から H で表現される．2.1 が特許で認められている 8 つの知識体系である．

クラス
: クラスとは各セクションを細分化し，セクションに 2 つの数字を付け加えたものである．
例として，A01 を挙げる．A がセクションであり，01 がクラスを表している．そしてセクションだけでは生活必需品という解説となるが，クラスが加わることで，A01 を農業；林業；畜産；狩猟；捕獲；漁業に関係する技術，発明であると解説しています．

表 2.1: セクション

	説明
A	生活必需品
B	処理操作
C	科学
D	繊維
E	固定構造物
F	機械工学
G	物理学
H	電気

サブクラス
　サブクラスとはクラスに 1 つの大文字を付け加えたものである．
　例として，A01B を挙げる．A01B は農業または林業における土作業：農業機械または器具の部品，細部または附属具一般を意味している．

メイングループ
　メイングループとはサブクラスに 1 つから 3 つの数字，斜線及び数字 00 を付け加えたものである．
　例として，A01B1/00 を挙げる．A01B1/00 は手作業具を意味している．

サブグループ
　サブグループとはメイングループにおける斜線部分の 00 がそれ以外の 2 つの数字になっているものである．
　例として，A01B1/02 を挙げる．A01B1/02 は鋤；ショベルを意味している．

これら 5 階層で表現する記号が組み合わされたものを完全分類記号と呼ぶ．IPC の構造の例として先程あげた A01B1/02 の図 2.2 を紹介する．この図は IPC の構造である A01B1/02 の構造を説明している．

2.3.2 File Index(FI)

FI は日本国特許庁が日本国内の特許にのみ適用した特許分類である．しかし，IPC の技術分類まま国内に提出される特許を分類すると，多量の特許がある IPC の特許分類に集中していまう．そのため，IPC を使用した検索が効率的に行われない．そこで，FI は IPC の利用を円滑に手段として，IPC をさらに展開した索引であり，展開記号及び，又は分冊識別記号という新たな記号を IPC に付加する形で作成された [23]．FI は IPC の完全

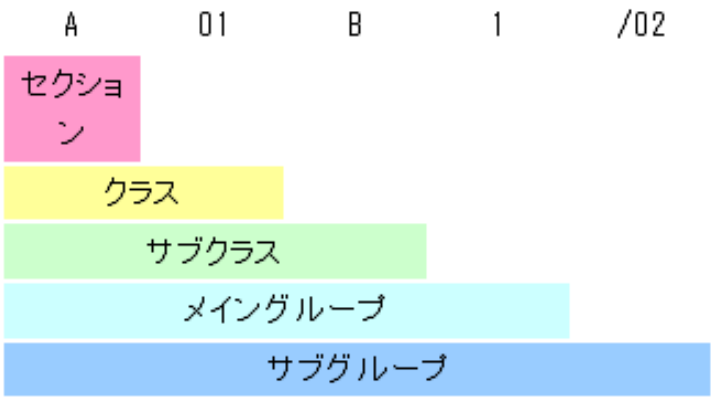

図 2.2: A01B1/02 の構造

表 2.2: FI 表記

FI 記号の表記形式	例示
IPC 記号	A21D 2/04
IPC 記号 + 分冊識別記号	B01D 53/02 B
IPC 記号 + 展開記号	B31B 1/00 301
IPC 記号 + 展開記号 + 分冊識別記号	C04B 35/58 104 B

分類記号に 3 桁の数字 (これは展開記号と呼ばれている) および，または 1 桁の大文字アルファベット (これは分冊識別記号と呼ばれている) を付け加えたものである．現在，FI は約 19 万の項目から構成されている．表 2.2 は FI の表記についての説明をしている [23].

IPC と FI のの細分化の概念をより分かりやすく示すため，具体例として，IPC 第 6 版の G11B20/18 と，FI の G11B20/18,542D との関係を以下図 2.3 に示します [18].

このように IPC をより細分化することで，効率的な特許の検索を可能にし，検索結果の内容をより絞り込むことが可能になる．

2.3.3 File Forming Term(F タ ー ム)

File Forming Term(F ターム) は，IPC や FI と共存し，かつコンピュータを利用した検索に適した新たな分類として導入された．この F タームは IPC の IPC や FI のように発明の技術内容，特徴によって分類されておらず，IPC や FI の特許分類とは異なる複数の技術的観点 (発明の目的，用途，材料，) によって分類することができる特徴を持っている [23, 29].

FI のみでは区分けが粗い分野もあり，近年発展した技術分野においては FI の表記におい

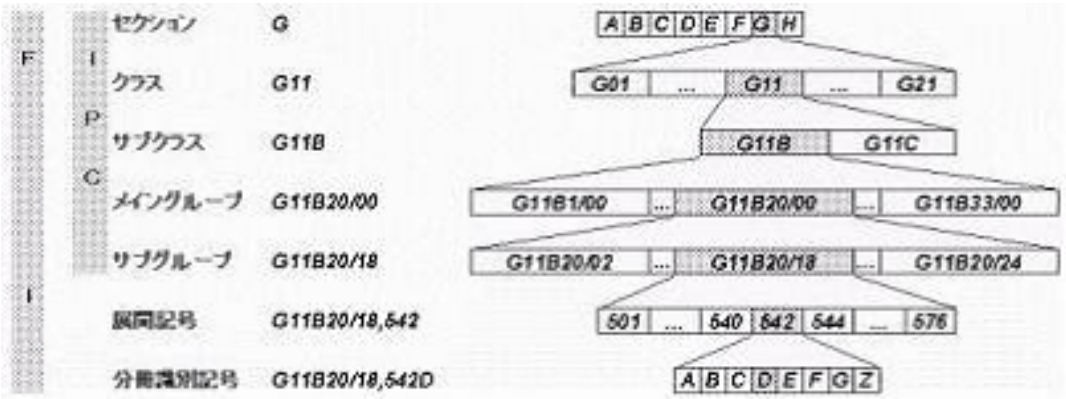

図 2.3: IPC と FI の関係

ても広範囲になり，そのため先行技術文献の調査において数多くの文献調査しなければならないことになる．これらの問題に対して FI を技術分野ごとに技術観点から細区分したものが F タームであり，多観点での解析，付与が可能であることが特徴である [23].
F タームは，FI で定められる一定の技術範囲ごとに区分して整備されており，区分された各技術範囲を「テーマ」と呼ばれている．この範囲のことをそのテーマの「FI カバー範囲」と呼ばれている [23]．各テーマは，その技術分野を端的に表現した「テーマ名」，英数字 5 桁のコードで構成されている「テーマコード」がある．現在，全技術分野が，約 2600 のテーマにより区分されており，そのうちの約 7 割に当たる 1800 のテーマにおいて F タームが作成されている．テーマがどのように表記されているかを表 2.3 を紹介する．

次に F タームについて詳しく説明をしていく．F タームとは,「テーマコード (英数字) 5 桁」 + 「観点 (アルファベット) 2 文字」+ 「数字 2 件」で構成されている.「観点」とは,その下に展開される複数の F タームを取りまとめる概念に対応して設定されるものである [23]．例として目的，機能，構造，材料，用途，製造方法，などがある．通常，観点と数字を合わせた 4 桁をさして「F ターム」と呼ぶ．図??と図??では F タームの表記の例と，特許庁で公開している特許電子図書館 (IPDL) で説明されている F タームの例を紹介する．

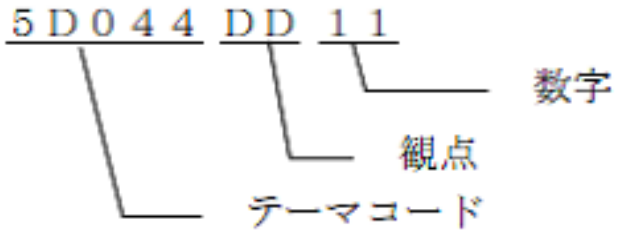

図 2.4: F タームの表記例

表 2.3: テーマコードのサンプル

テーマコード	テーマ名
4J001	ポリアミド
4J002	高分子組成物
4J004	接着テープ
4J005	ポリエーテル
4J006	ポリハロゲン化オレフィン
4J011	重合方法 (一般)
4J015	重合触媒
4J016	重合後の処理
4J017	後処理による化学的変性
4J019	オレフィン系重合体
4J020	スチレン系重合体
4J022	酸素含有重合体
4J023	アクリル系重合体
4J024	不飽和特殊重合体
4J025	ジエン, アセチレン, 炭化水素, 他系重合体
4J026	グラフト, ブロック重合体
4J027	マクロモノマー系付加重合体
4J028	オレフィン, ジエン重合用触媒
4J029	ポリエステル, ポリカーボネート
4J030	硫黄, リン, 金属系主鎖ポリマー
4J031	炭素-炭素不飽和結合外反応のその他樹脂等
4J032	ポリオキシメチレン, 炭素-炭素結合重合体
4J033	フェノール樹脂, アミノ樹脂
4J034	ポリウレタン,　ポリ尿素
4J035	ケイ素重合体
4J036	エポキシ樹脂
4J037	顔料, カーボンブラック, 木材ステイン
4J038	塗料, 除去剤
4J039	インキ, 鉛筆の芯, クレヨン
4J040	接着剤, 接着方法
4J041	ゴムの処理
4J042	天然高分子誘導体, 膠, ゼラチン
4J043	含窒素連結基の形式による高分子化合物一 般
4J100	付加系重合体, 後処理, 化学変成
4J127	マクロモノマー系 付加重合体
4J128	付加重合用遷移金属, 有機金属複合 触媒
4J200	生分解性ポリマー
4J246	けい素重合体
4J777	コンビナトリアルライブラ リ (他の高分子)

5D044	デジタル記録再生の信号処理	情報記録
	G11B20/10-20/16,351@Z	

観点		Fターム										FI適用範囲
AB	AB00	AB01	AB02	AB03		AB05	AB06	AB07	AB08	AB09	AB10	G11B20/10-20/16,351@Z
	記録情報	・コンピュータ情報	・・プログラム	・・データ構造に特徴のあるもの		・オーディオ	・・コード化オーディオデータ（MIDI等）	・ビデオ	・・静止画	・・文字情報	・その他＊	
BC	BC00	BC01	BC02	BC03	BC04	BC05	BC06		BC08		BC10	
	記録再生原理	・磁気	・光	・・再生専用	・・記録再生	・・・追記	・・・書換え（光磁気等）		・各方式の混在型		・その他＊	
CC	CC00	CC01	CC02	CC03	CC04	CC05	CC06	CC07	CC08	CC09	CC10	
	記録担体，トラック	・テープ	・・長手方向に平行なトラックを持つもの	・・長手方向に平行でないトラックを持つもの	・ディスク	・・同心円トラックを持つもの	・・スパイラルトラックを持つもの	・ドラム	・カード	・複数の記録担体	・その他　＊	
DE	DE00	DE01	DE02	DE03	DE04							G11B20/10; 20/10@A- 20/10@Z; 20/10,301; 20/10,301@A; 20/10,301@B; 20/10,301@Z; 20/10,311; 20/10,321; 20/10,321@A;
	記録担体上の記録信号の配列	・情報全体の構成，配置	・・フォーマッティング（初期化）	・・情報全体のブロック又はセクタ分割	・・フォーマット変換							
		DE11	DE12	DE13	DE14	DE15		DE17	DE18	DE19		
		・主情報およびその配置（一次情報と二次情報）	・・主情報のブロック又はセクタ分割	・・チャンネル分割	・・マルチメディア（複数の一次情報を記録）	・・・複数の異なる方式で記録（通常TVとHD等）		・・一次情報に付随する二次情報	・・・字幕情報	・・・高速再生用情報		
			DE22	DE23	DE24	DE25		DE27	DE28	DE29		

図 2.5: F5タームの表

2.4 特許自動分類手法に関する先行研究

特許自動分類とは，計算機システム等を用いて，特許の発明の特徴から特許分類体系の中から的確な分類を選択し，その分類記号を付与することである．この研究に関してはNTCIR において数多くの研究がなされている．その研究の中において F タームごとに自動で分類する研究が行われている [5, 6]．これらの研究では，F タームが持っている「観点」の特徴に注目し，F タームを利用した特許の検索は非常に有効であると考えられており，特許審査官が特許を審査する際にも使用され，特許出願時にする先行調査においても使用されている．そこで，特許の情報から F タームを的確に自動的に付与できるならば，特許出願における問題の一つである先行技術の調査にかかる時間短縮，費用の縮小が図れるであろうと考えられる．この研究において主な手法が F タームを利用した分類である．これは各 F タームごとに各々の特許がある F タームの特徴に当てはまるかどうかを出力することで，各々の特許にその F タームが正しく付与されているかるかどうかの判定を行う方法である．この研究の構成を図 2.6 に示す．この分類の手法で SVM を利用した分類の研究が行われている [8, 9, 7]．F タームの特許自動分類おいては，教師つき学習と呼ばれる訓練データに正解と不正解の情報を与えることで正解のパターンと不正解のパターンを導くことで，テストデータを与えた場合にパターンからテストデータが正解か不正解かを出力する．F タームによる特許自動分類においてはテストデータの正解を元々特許につけられている F タームを正解とし，その出力された結果を比較することで，精度をはかる．その精度をはかる尺度として F 値が主に使わている．F 値については，4 章の「特許自動分類の評価方法」の節で解説する．

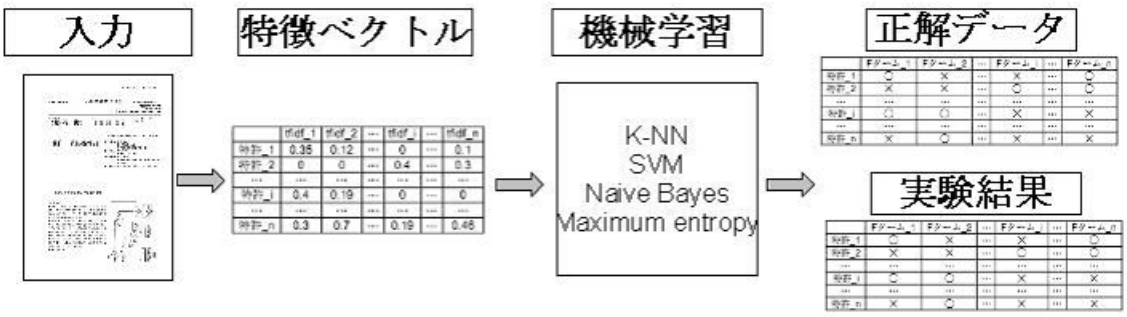

図 2.6: 先行研究の構成

第3章　カーネル手法

本章では，カーネル手法の特徴と定義，条件を解説する．また，カーネル手法はSVMと組み合わせて利用されることが多く，また，本稿においてもSVMと組み合わせて使うため，SVMの概要を説明し，カーネル手法との関係を述べる．また，テキスト分類に関する概要と，カーネル手法を利用する利点についても述べる．

3.1　カーネル手法とは

カーネル手法とは，データを特徴空間に写像することと，特徴空間において線形パターンを発見する学習アルゴリズムからなっている．データを特徴空間に写像することは，言い換えると非線形データや離散構造のデータ (vector, string, tree, graph, text, etc) を高次元に射影し，線形問題に置き換えることができる [?]，[10]，[25]．これにより今まで，統計学や機械学習において研究されてきた線形の関係を発見するアルゴリズムが十分に理解され，カーネル手法が機能することを証明している．また，カーネル手法はSVMなど多くの識別学習問題において応用され，機会学習，データ解析，文書分類など様々な分野にて注目を集めている [31]．この手法の特徴は入力空間におけるデータを新たな特徴空間へ写像することで分類や識別などのタスクを簡略化できることである．また，特徴空間で各点の内積で表現することで，最小の特徴集合を見定めることにより，高次元での写像時の計算能力と汎化能力のいずれも低下する，”次元の呪い”の現象を回避することを可能にした．このように各点の内積を計算することで，その関係は類似度という定義をすることが可能になり，これは入力空間上での各点は類似度という尺度により，似ているか，似ていないかを直感的に判断できること意味している [10, 26]．このような特徴からカーネル手法は画像識別，バイオインフォマティックス，離散構造に対する分類などの様々な分野で応用されている．

3.1.1　カーネル関数の定義・条件

定義：カーネル関数

入力空間における X, 特徴空間を F としたとき，それらの特徴空間への写像は

$\Phi : X \to F$ となる．さらに2点 $x_i, x_j \in X$ が与えられたとき，それらの特徴空間において写像点 $\Phi(x_i), (x_j)$ と表し，カーネル関数を用いて $K = \langle \Phi(x_i), \Phi(x_j) \rangle$ と定義される．

カーネル関数が特徴空間にて確実に内積であることを保証するための条件として,
1. 関数の対称性

$$K(\mathbf{x},\mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle = \langle \Phi(\mathbf{y}), \Phi(\mathbf{x}) \rangle = K(\mathbf{y},\mathbf{x}) \tag{3.1}$$

2. コーシーシュワルツの不等式

$$K(\mathbf{x},\mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle^2 \leq \| \Phi(\mathbf{x}) \|^2 \| \Phi(\mathbf{y}) \|^2 = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}) \rangle \langle \Phi(\mathbf{y}), \Phi(\mathbf{y}) \rangle = K(\mathbf{x},\mathbf{y}) \tag{3.2}$$

3. マーセルの定理
X は $\Re$ の部分集合で，関数 $K : \langle X \times X \rangle \to \Re$ は連続かつ対象 $K(\mathbf{x},\mathbf{y}) = K(\mathbf{y},\mathbf{x})$ とする．このとき，関数 K が様収束する級数：

$$K(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{\infty} a_j \phi_j(\mathbf{x}) \phi_j(\mathbf{y}), a_j > 0 \tag{3.3}$$

によって展開可能となる必要十分条件は

$$\int_{\mathbf{x}\times\mathbf{x}} K(\mathbf{x},\mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0, \forall f \in L_2(\mathbf{X}) \tag{3.4}$$

である．これは $L_2(\mathbf{X})$ の任意の有限部分集合に対して対応する行列が半正定値となることを意味している．
以上 3 つの条件を満たす必要がある．
これらの条件を満たす代表的なカーネル関数を示す.

- 線形カーネル：$K(\mathbf{x},\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$
- 多項式カーネル：$K(\mathbf{x},\mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})$
- ガウシアンカーネル：$K(\mathbf{x},\mathbf{y}) = \exp -\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}$
- シグモイドカーネル： $K(\mathbf{x},\mathbf{y}) = \tanh k(\mathbf{x} \cdot \mathbf{y}) - \theta$

一般的に，カーネル関数をデータに適用する際に考えることは，カーネル関数がそのデータに適切な関数であるかである．この際，カーネル関数の選択は過去の他問題においてどのような関数が適用されていたかを参考にし選択するといったヒューリスティックなものになる．

3.1.2 カーネル行列

カーネル関数によって求められた行列は，一般にカーネル行列，または特徴空間内ではグラム行列と呼ばれ，入力空間における各データ間の内積を評価するものである．さらに写像空間におけるすべての点の位相位置を完全に決定付けるものである．また，その条件として関数と同様に行列が対称性 $K_{ij} = K_{ji}$，半正定値性を満たすことが挙げられる．カーネル行列は

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \vdots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{bmatrix} \tag{3.5}$$

と表す．の行列は対称性，半正定値性を満たしていれば，SVM による学習分類が可能になる．つまり，非線形データや離散構造のデータをカーネル関数により計算し，作成したカーネル行列が対称性，半正定値性を満たすならば，SVM による学習分類が可能になるのである．

3.2 Support Vector Machine

Support Vector Machine (SVM) は，高次元特徴空間において線形関数の仮説空間を用いる学習システムである．SVM は Vanpnik らによって導入された学習法で，教師付き学習の一つであり，現在知らている手法の中で高認識率を達成することができる手法である．また，カーネル関数を取り扱うことのできる学習器である．本稿ではまず SVM の学習アルゴリズムについて説明する．SVM の基本的概念は，トレーニング集合からマージン最大化を満たす線形しきい素子を学習し二値分類を構成するものである．今，入力ベクトル $(x_i, y_i), i = 1, ..., n, x_i \in \Re^n, y_i \in \pm 1$ で構成されたトレーニング集合 S が与えられたとして，この集合を二値に識別する際の決定境界は超平面，

$$\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0, [\mathbf{x}] \in \Re^{\mathbf{n}} \tag{3.6}$$

で表せる．ここで示された $\mathbf{w}$ は超平面からの重みベクトル，b はバイアス，しきい値である．線形しきい素子は入力ベクトルより識別関数 $f(\mathbf{x})$

$$f(\mathbf{x}) = sgn(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) \tag{3.7}$$

となる 2 値の出力を計算する．ただし $(\mathbf{w}, b) \in \Re^n \times \Re$ は関数を制御するパラメータである．この式は，$f(\mathbf{x}) > 0 (f(\mathbf{x}) < 0)$ の場合，入力点である $\mathbf{x}$ は正また負として識別されることを意味する．さらに，線形分離では式 3.6 で表される超平面が無数にあることから，これを図 3.1 で示すような超平面のうち一番近いサンプル点までの距離，マージンを導入する．一般的に幾何マージンは，入力空間での決定境界から入力点までの距離に値する．ここで，識別関数に幾何マージンを設けた場合，超平面 $(\mathbf{w}, b)$ に対する入力ベクトル (x_i, y_i) の幾何マージンは，

$$\gamma_i | \frac{y_i((\mathbf{w} \cdot x_i) + b)}{\| \mathbf{w} \|} \tag{3.8}$$

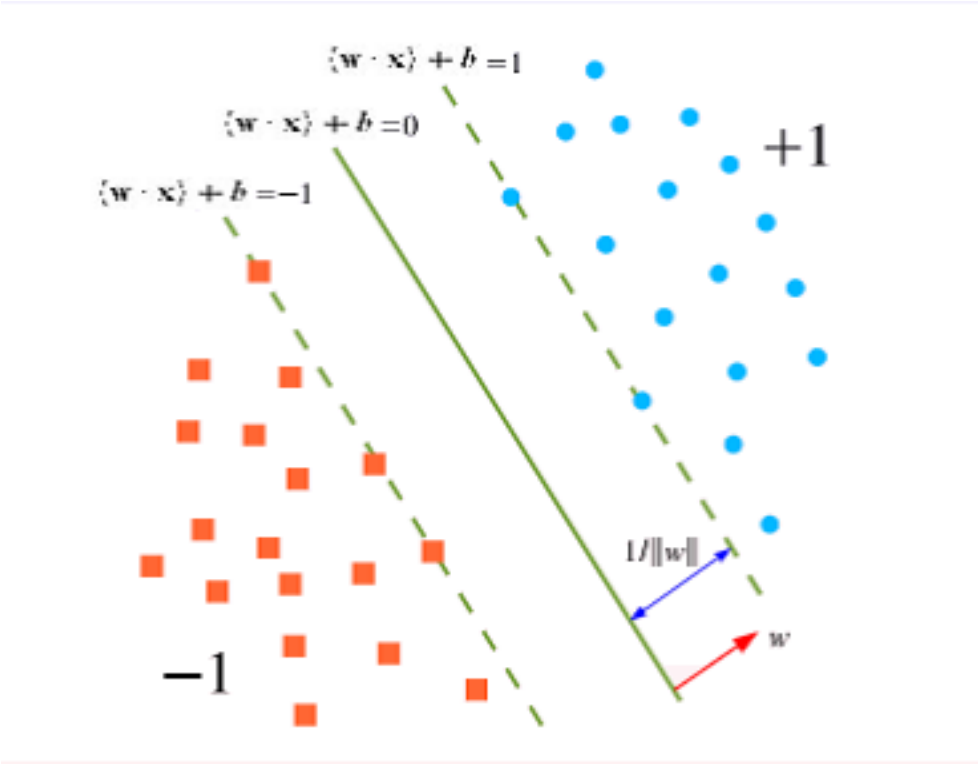

図 3.1: マージンの概略図

式 3.8 関数マージン $y_i((\mathbf{w} \cdot x_i) + b)$ を正規化により $\| \mathbf{w} \|$ で割り当てられた幾何マージン γ_i は $\gamma_i > 0$ の場合，入力ベクトルを識別できることに値する．
SVM の分類の目的の一つは，上記のトレーニング集合 S より形成される空間で分離する最適な超平面を学習することであり，最適化問題を解くことである．最適化問題として，線形分離可能な場合の最大マージンクラス分類がある．最大マージンクラスとはトレーニング集合でのすべての超平面における最大幾何マージンを求めるものである．最大マージ

ンは点の少ない部分集合に極めて依存してしまうことでハードマージンとも呼ばれている．これはマージン制限下で最適化問題を解決することにより超平面をもとめるものである．これに対し，図 3.2 では誤差を表すスラック変数 ξ を導入したソフトマージンによる最適化問題も存在する．以下，ハードマージン，ソフトマージンについて説明する．

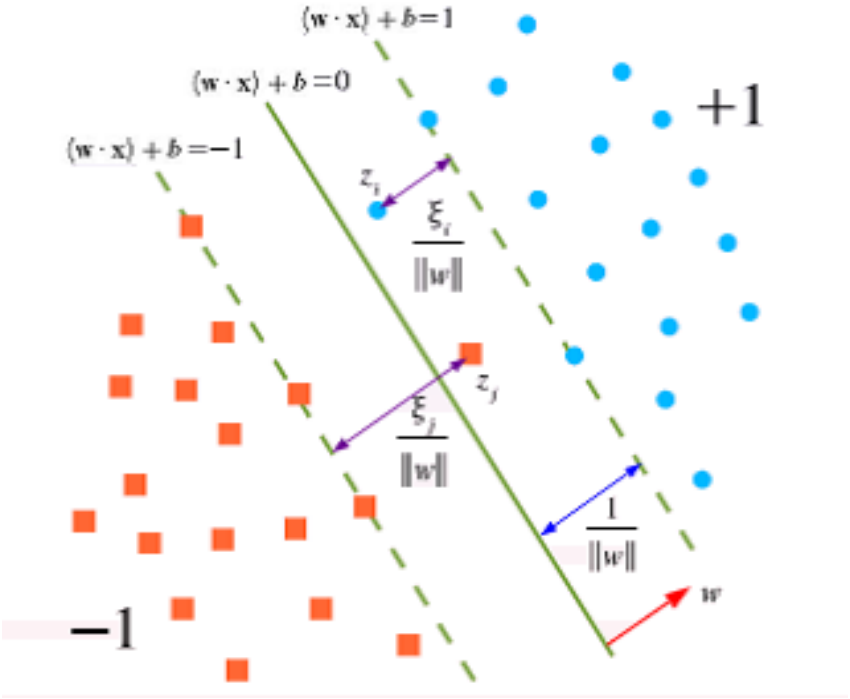

図 3.2: ソフトマージンの図

ハードマージンによる最適化

ラベル付された集合 $S = (x_1, y_1), ..., (x_n, y_n)$ が与えられたとき，最適化問題を解決する超平面 $(\mathbf{w}, b)$ は，

$$\begin{aligned} &\min_{\mathbf{w},b}\langle \mathbf{w}\rangle, \\ &\text{subject to } y_i(\langle \mathbf{w}\cdot \mathbf{x}_i\rangle + b) \geq 1, i = 1, ..., n \end{aligned} \tag{3.9}$$

となる．式 3.9 における $\mathbf{w}$ の最適解を $\mathbf{w}^*$ とした場合の幾何マージン $\gamma = \frac{1}{\|\mathbf{w}^*\|_2}$ をもつ最大マージンクラス分類問題を実現する．この最適化題を，ラグランジュ定数を導入し，双対問題へ変換することで，

$$\begin{aligned} &\max W(\alpha) = \textstyle\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (x_{i,x_j}) \\ &\text{subject to } \textstyle\sum_{i=1}^{n} y_i \alpha_i = 0, \alpha_i \geq 0, i = 1, ..., n \end{aligned} \tag{3.10}$$

が得られる．パラメータ α^* は式 3.10 の最適化問題の最適解になる．これにより最適な重みベクトル $\mathbf{w}^* = \quad sum_{i=1}^{n} y_i \alpha_i^* x_i$ となり幾何マージンが $\gamma = \frac{1}{\|\mathbf{w}^*\|_2}$ の最大マージ超平面を実現する．このハードマージン開放はラベル付されたトレーニング集合を線形分離できる．つまり，線形識別のトレーニング誤差がゼロの完全なトレーニング集合が存在する時のみ用いることが可能で，実世界のようなデータにノイズを含む場合の線形分離に対しては，多少の識別誤りがあってもよいという制約が必要である．

ソフトマージンによる最適化

線形分離において多少の識別誤りがあってもよいという考えを一般的にソフトマージンと呼ばれている．ソフトマージンという考えでは，式??に対して図 3.2 で表すようなスラック変数 ξ を導入し，

$$\begin{gathered} y_i((\mathbf{w}, \mathbf{x}_i) + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, i = 1, ..., n \end{gathered} \tag{3.11}$$

と変換する．このような制約下で最適化解を求めることがソフトマージンによる最適化である．
ソフトマージンによる最適化を説明する．
1-ノルムソフトマージンによる最適化ではハードマージンによる最適化と同様に，ラベル付された集合 $S = (x_1, y_1), ..., (x_n, y_n)$ が与えられたとき，最適化問題を解決する超平面 $(\mathbf{w}, b)$ は，

$$\begin{gathered} \min_{\mathbf{w},b} \langle \mathbf{w} \rangle + C \sum_{i=1}^{n} \xi_i, \\ \text{subject to } y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, i = 1, ..., n \\ \xi_i \geq 0, i = 1, ..., n \end{gathered} \tag{3.12}$$

となる式 3.12 を $\mathbf{w}$ の最適解を $\mathbf{w}^*$ とした場合の幾何マージン $\gamma = \frac{1}{\|\mathbf{w}^*\|_2}$ をもつ 1-ノルムソフトマージン分類問題を解くことを実現する．これも，ハードマージンと同様，式 (3.12) に相当する双対問題は，

$$\begin{gathered} \max W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (x_{i,x_j}) \\ \text{subject to } \sum_{i=1}^{n} y_i \alpha_i = 0, C \geq \alpha_i \geq 0, i = 1, ..., n \end{gathered} \tag{3.13}$$

が得られる．パラメータ α^* は式 (3.13) の最適化問題の最適解になる．

2-ノルムソフトマージンによる最適化ではハードマージンによる最適化と同様に，ラベル付された集合 $S = (x_1, y_1), ..., (x_n, y_n)$ が与えられたとき，最適化問題を解決する超平面 $(\mathbf{w}, b)$ は，

$$\begin{gathered} \min_{\mathbf{w},b} \langle \mathbf{w} \rangle + C \sum_{i=1}^{n} \xi_i^2, \\ \text{subject to } y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, i = 1, ..., n \\ \xi_i \geq 0, i = 1, ..., n \end{gathered} \tag{3.14}$$

となる式 (3.14) を $\mathbf{w}$ の最適解を $\mathbf{w}^*$ とした場合の幾何マージン $\gamma = \frac{1}{\|\mathbf{w}^*\|_2}$ をもつ 1-ノルムソフトマージン分類問題を解くことを実現する．これも，ハードマージンと同様，式 3.14 に相当する双対問題は，

$$\begin{array}{c} \max W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \left((x_{i,x_j}) + \frac{1}{C} \xi_{ij} \right) \\ \text{subject to } \sum_{i=1}^{n} y_i \alpha_i = 0, C \geq \alpha_i \geq 0, i = 1, ..., n \end{array} \tag{3.15}$$

が得られる．パラメータ α^* は式 (3.15) の最適化問題の最適解になる．

3.2.1 SVM におけるカーネル

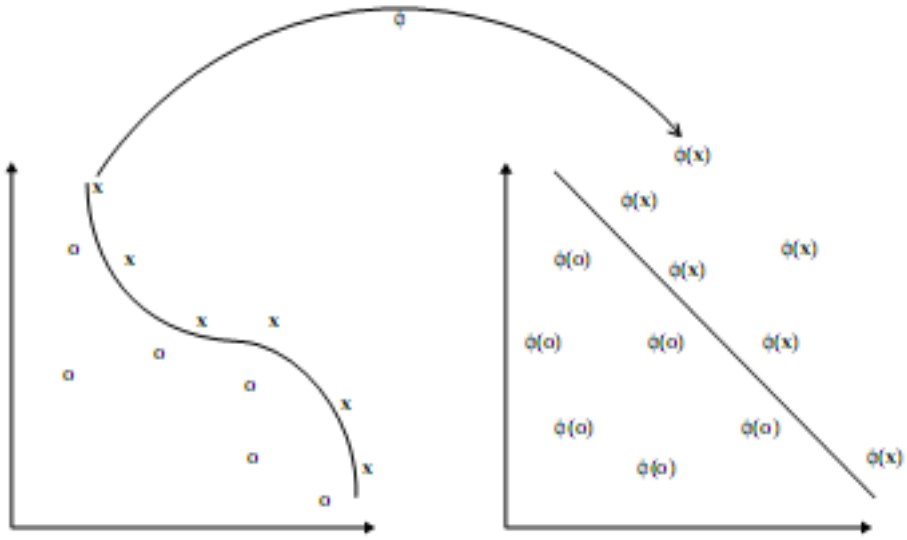

図 3.3: 特徴空間への写像

識別問題では実世界で与えられるデータは線形であるものは珍しく，非線形における分類では元の空間上で分類できない場合が多く，そのような問題にソフトマージンを適用「したにしても，非線形で複雑な識別問題に対して精度の良い適切な構成をするのは困難である．そこで，そのような元の空間における非線形のデータに対して，特徴空間にデータを写像し，その空間において線形分離するのである．そのような場合，考慮する仮説集合は，

$$f(\mathbf{x}) = \sum_{i=1}^{n} w_i \phi_i(\mathbf{x}_i) + b \tag{3.16}$$

という関数となる．線形関数による学習がデータをグラム行列の要素を通じてのみ現れる双対問題として扱える．そのため，決定規則が入力データの内積のみで評価できるように

なり，

$$f(\mathbf{x}) = \sum_{i=1}^{n} \langle \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) \rangle + b \tag{3.17}$$

となる．ここで元の空間から高次元特徴空間への写像において元の空間における入力データ同士の距離関係を維持するために，カーネル関数の定義式 $K = \langle \Phi(x_i) \cdot \Phi(x_j) \rangle$ を式 (3.17) に代入すると，

$$f(\mathbf{x}) = \sum_{i=1}^{n} K(\mathbf{x}_i) \cdot (\mathbf{x}) + b \tag{3.18}$$

という式が得られる．すなわち，SVM 上での符号関数は，

$$f(\mathbf{x}) = sgn\left(\sum_{i=1}^{n} K(\mathbf{x}_i) \cdot (\mathbf{x}) + b \right) \tag{3.19}$$

となる．これにより特徴空間内での学習に必要な写像点 $\phi(x)$ を知る必要がない．また，カーネル関数を 1-ノルムソフトマージン，2-ノルムソフトマージンに代入することも可能である．
以上のように元の空間で線形分離できない場合，高次元特徴空間への非線形写像により，その特徴空間において線形分離を行う解決策をカーネル関数は実現し，特徴空間上での線形関数の学習を可能にする．

3.3 テキスト分類におけるカーネル手法について

近年，デジタル化されたテキストが増大し続けているため，それらのデータを人手で分類することはほぼ不可能となっている．この結果，自然言語のテキスト文書を自動的に分類することが，研究の対象となり，人工知能や機会学習，自然言語処理など様々分野において盛んに研究が行われている [12]．このような研究において，情報検索の研究者がテキスト情報を表現する共通の方法は，ベクトル空間モデル (vector space model) である．このベクトル空間モデルを用いてカーネルを作成することにより、カーネル手法をテキスト分類に応用できるのである [10]，[?]．この結果、カーネル手法が幅広いタスク，例えば，相関分析，クラス分類，ランク付け，クラスタリングなど，に適用可能になり，これらのタスクはテキスト分類においては，文書分類，フィルタリングとして名前を変えて登場することになる [12].

3.3.1 テキスト分類について

テキスト分類の目的は様々である．文書分類のタスクはニュースを政治，経済，スポーツなどのカテゴリに自動で分類することや，電子メールのフィルタリングなどが行われて

いる [10]，[19]．テキスト分類において，使われているのが SVM を利用した分類である．SVM を利用したテキスト分類は，テキスト分類の研究において有効な方法であるという結果を得ている [19]，[10]. しかし，SVM は性質上様々なパラメーターをテキスト情報の特徴に合わせて調整を行わなければならず，さらには，テキスト情報の特徴量が増えるとそれに合わせて計算量も指数関数的に増大するなどの弱点もある．

3.3.2 テキスト分類におけるカーネル法の有効性について

カーネル手法をテキスト分類に適用「できるのは，テキスト情報の特徴をベクトル空間モデルに変換し，そのベクトル空間モデルにカーネル手法を適用することで可能となる．ベクトル空間モデルの詳細については 4 章で解説する．テキスト文書をベクトル空間モデルに変換することでカーネル手法のアルゴリズムを多岐に適用できるようになり，これにより，SVM のみを利用するよりも適したアルゴリズムと少ない計算量で分類を行うことが可能になるという利点がある [10, 12].

第4章　カーネル手法を適用した特許の分類

日本国における特許出願から審査までの期間が長いことに問題を抱えており，それに対応するために特許の審査官を増員し，徐々にではあるが処理能力の向上しつつあるが，未だに審査されるまでの期間が長く，審査されていない特許が多数存在する．この問題に対して計算機を利用した特許についての処理の研究が行われるようになっている．その中でも特許に付与されているIPCやFI,Fタームを利用した分類の研究が行われている．そのような中で，本稿では，テキスト情報の分類に用いられるカーネル手法を特許自動分類の研究に適応し，最適なモデルの探す手法に解説していく．

4.1　特許自動分類のカーネル手法の適用

近年，デジタル化されたテキスト情報が人手で分類されることが不可能なほど増大したことで，テキスト情報の分類の研究が盛んに行われ，その効果は評価されている．本稿では，特許自動分類をテキスト情報の分類の一領域と捉え，特許というデジタル化されたテキストをテキスト分類で用いられているカーネル手法を適用することで，今まで研究されてきた特許の分類の方法との比較を行う．そして，特許における2つのカーネル法の最も最適なモデルを発見することで，特許の分類におけるカーネル手法の効果を示すことを目標とする．

この目標に対し，カーネル手法を適用し，各Fタームごとにテストデータを正負に分類を行い，その正負の情報と，元々テストデータに与えられているFタームとを比べることで，カーネル手法を適用した場合での精度の違いを観察し，カーネル手法の最適な設定を求めることで精度の向上を図る．先行研究において，SVMが一般的に使われており、その精度も他の分類器を利用した場合より高い精度で分類している [7]，[9]．特許には他の文書やニュースに比べラベルが多く，ひとつの特許に複数のラベルがある多値分類という分類の手法を利用している．そこで先行研究ではこれらの問題に対処するためにSVMを特許の分類に最適なアルゴリズムを既存のアルゴリズムを改良，または，正負の偏りを解消する研究に重点が置かれていた．しかしながら，どの研究においても改善が見られてはいるが，改善の余地が残されている．そのため，本稿では，焦点を変え，特許の情報の殆どを占める，文書情報についてに注目した．文書分類の先行研究ではカーネル手法を用いた分類方法において良い精度が得られている [13]．カーネル関数を適用する際には，最適

なパラメーターを得ることでパフォーマンスが向上することから，本稿では，カーネル関数の適用の際の最適なパラメーターを求めることを目標とする．テキスト情報にカーネル手法を適用する際に必要となるベクトル空間モデルについての解説を行い，次にテキスト情報からベクトル空間モデルに変換する際に欠損する情報を少なくする手法である $tf \times idf$ について解説をし、ベクトル空間モデルがカーネル手法で扱えることを解説し，次節以降では本稿で利用するカーネル関数の解説，多値分類の手法の解説，不均衡データに対する対応を解説していく．

4.2 特許データのベクトル空間モデル化

テキストをカーネル手法で扱うにはテキストをベクトル空間モデルに置き換える必要がある．このベクトル空間モデルで比較的単純で，一般に使われているモデルが，単語集合 (bag-of-words) である．この単語集合においては文書内の単語のみに焦点を当てている．これは，単語の順序に関する情報がないので、文法情報が失われ，また，2 個や 3 個の単語によって正確な意味ができる名詞句などの句は 1 つ 1 つの単語に分解されるため，句の意味を失うことがある [12]．図 4.1 は特許文章をベクトル空間モデルに変換し，その後，カーネルマトリックスに変換し，SVM の学習器で学習するまでのプロセスを表している．

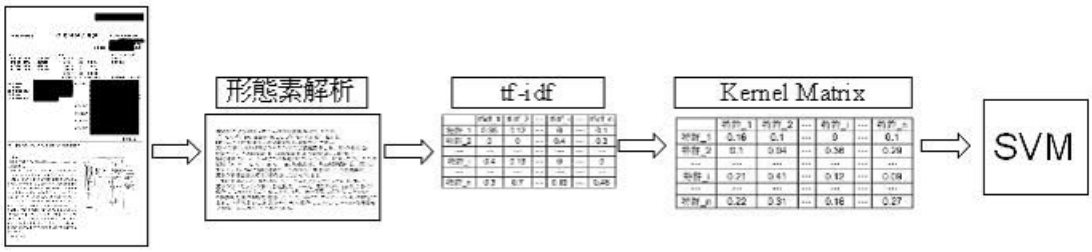

図 4.1: ベクトル空間モデル化のプロセス

4.2.1 単語集合 (bag-of-words)

単語集合とは，文書 d は，用語辞書からの用語を添字とし,「対応する用語が存在するか否かの変数」を値とするベクトル $\phi(d)$,

$$\phi : d \longmapsto \phi(d) = ((tf(tf_1, d), (tf(tf_2, d), ..., (tf(tf_N, d),)) \in \Re^N \tag{4.1}$$

と表現できる [10, 12]．$tf(t_i, d)$ は文書 d_j の単語 t_i の頻度とする．これにより，テキスト情報は次元 N の空間へ写像される．一般的にこの空間の次元は非常に大きな数字になる．この単語集合の欠点は単語の順序や，文章が持っている文法についての情報など，文脈や

言葉としての意味などの情報が失われることにある．この問題に対して，単語に重要度を設定したベクトル空間モデルを導くことを行う．

4.2.2 $tf \times idf$

前節で述べたように，単語集合においてすべての単語が重要性を持っているわけでない．そこで単語に対して重みをつけることで，各々の単語に重要性を重視した関数を加えることで，単語に意味を持たせることで，元のテキスト情報の欠損情報を少なくでき，より正確な分類を導くことができる [10, 12].
その方法として，idf という計算方法がある．idf は単語を文書頻度の逆数 (inverse document frequency) の関数として重み付ける．l 個の文書があるとき，$df(t)$ を単語 t を含む文書の数とすると，単語 t に対する idf は，

$$idf(t) = ln\left(\frac{l}{df(t)}\right) \tag{4.2}$$

と与えられる．そして，d_n における $tf \times idf$ は次のように表すことができる．

$$\phi_n(d) = [tfidf(t_i, d_n), tfidf(t_2, d_n), ..., tfidf(t_N, d_n)] \in \Re^N \tag{4.3}$$

ただし，tf_i を文書 d_n での項目 i の発生数，idf_i を総文書数と項目を含む文書数の比率とする

4.3 カーネル手法の適用

前節で表現された $tf \times idf$ はベクトル空間モデルとして定義できる．それにより，関係するカーネル手法は，

$$K(d_1, d_2) = \langle \phi(d_1) \cdot \phi(d_2) \rangle = \sum_{j=1}^{N} tfidf(t_j, d1) tfidf(t_j, d_2) \tag{4.4}$$

となり，この関数は陽に構築した特徴空間での内積であるから正当なカーネルである．したがって，このカーネル行列は常に半正定値となり，クラス分類に $K(d_1, d_2)$ や，他のカーネル関数を利用して SVM を使用できる [10, 12].
このことから，テキスト情報の一種である特許文書をベクトル空間モデルとして表現し，カーネル手法を適用できる．本稿では，適用するカーネル関数を，線形カーネルと RBF カーネルの 2 種類に限定し、そのパラメーターのチューニングをすることで最も精度の高いモデルを探る．この 2 つのカーネル関数は，他の研究においても一般的に利用され，ヒューリスティックにチューニングすることで最適なパラメーターを取得することで，よい評価を得ている [10]．また，図 4.2 は本稿における特許データからベクトル空間モデル

を生成し，その後，従来と違いカーネル手法を適応して後にSVMで分類するプロセスを表現している．カーネル手法を適応することで，従来の研究の様にSVMのアルゴリズムの向上を行うより的確なアルゴリズムを選択することで分類の精度の向上が見込めるプロセスを踏むことでカーネル手法の可能性と特許の自動分類の精度の向上を図る．

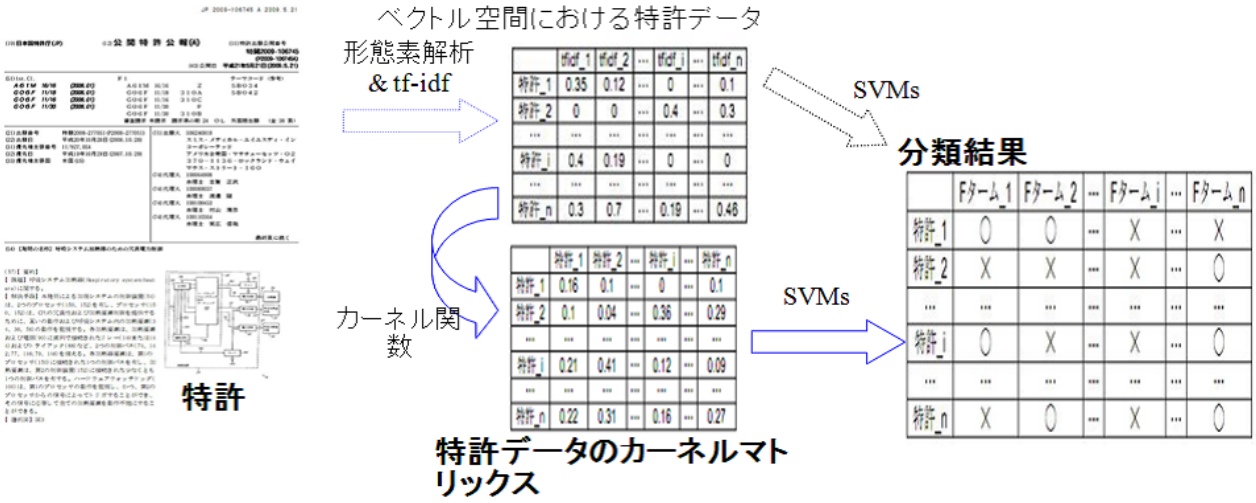

図 4.2: One-vs.-rest 法

4.3.1 線形カーネル

線形カーネルは3章で示したように以下の式で表現される．

$$K(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y} \tag{4.5}$$

この関数はSVMのソフトマージンを利用しているため，ソフトマージンの式におけるペナルティ項の「C」パラメーターを設定することで，最適なモデルを得ることができる．

4.3.2 RBF カーネル (Radail Basis Function)

RBF カーネルは以下の式で表現される．

$$K(\mathbf{x}, \mathbf{y}) = \exp - \frac{\| \mathbf{x} - \mathbf{y} \|^2}{\gamma} \tag{4.6}$$

RBF カーネルはカーネル手法の中で一般的に利用されるカーネル関数であるが，その性能を引き出すにはRBF カーネルが持っている2つのパラメーターを最適な設定にすることが必要である [32]．2つのパラメーターは，一つはソフトマージンの式のペナルティ項である「C」パラメーターであり，もう一つがRBF カーネルの式の「γ 」パラメーターである．

4.4 多値分類に対する工夫

もともと，SVM は 2 値分類を行う分類器である．しかし，1 つの特許にいくつもの F タームが付与さているため，2 値分類を拡張し，多値分類を適用する必要性がある．これに対して，2 つの手法があり，一つが One-vs.-rest 法であり，もう一つが Pairwise 法である．One-vs.-rest 法は k 個のクラスに対し，ある一つのクラスであるか，それ以外であるか，に分類する手法である．一方，Pairwise 法は k 個のクラスから，任意の 2 つのクラスを選び，それに関する 2 値分類の分類器を ${}_nC_k$ 個構築する手法である [27].
本稿では，先行研究において一般的に使われている One-vs.-rest 法を使って，特許の分類を行っていく．図 4.3 は特許自動分類に対する One-vs.-rest 法を表現している．各特許に $tf \times ids$ から得られたベクトル空間モデルとそれぞれが持っている F タームから，F ターム$_1$ と，それ以外の F タームをもっている，という 2 値分類の訓練データを作成し，テストデータも同様に作成し，それを F ターム$_2$ と，それ以外の F ターム，F ターム$_i$ と，それ以外の F ターム，そして，F ターム$_n$ と，それ以外の F タームをもっている，というようにデータを作成する．その後，各々のデータセットを学習器で学習させ，テストデータで与えられる F タームを予測し，結果を得る．その自動で付与された F タームをテストデータの特許ごとにまとめ，その結果と元々テストデータに与えられている F タームとの比較を行うことで実験の精度を測るという方法を取る

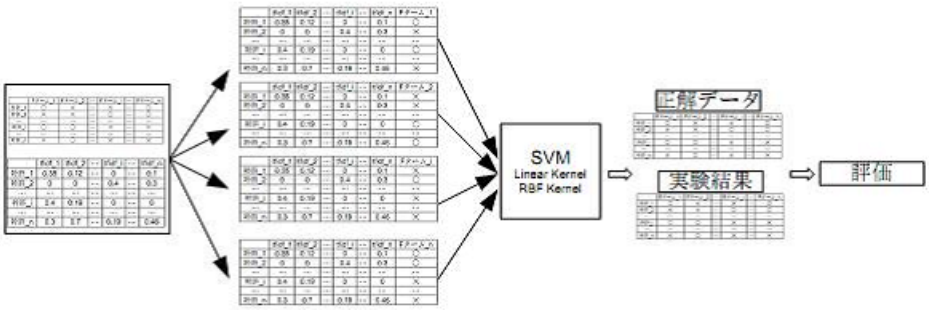

図 4.3: One-vs.-rest 法

4.5 特許自動分類におけるクラス不均衡

先行研究において，前節で解説した One-vs.-rest 法を利用した多値分類への工夫を行っているが，それぞれのクラスにおいて，クラスの不均衡が起こる問題がある．これは全データの内，ある F タームを持っている割合が 1 割など全体に対して少数であることである．これは，過学習という機会学習における一般的な問題であり，これを回避するために，本稿では，2 値の内どちらか一方に偏ったデータは実験に用いないことにする．そうすることで，自動分類の精度の低下を防ぐ．

第5章　実験と結果

本章では，先行研究における実験方法，実験に使用するデータ，実験の前準備，実験環境，実験方法，実験結果，とそれぞれについて解説を行う．

5.1　実験の目的

本稿では，4 章で解説したように，テキスト情報にカーネル手法を適用し，その際に適用するカーネル関数のパラメーターをチューニングし，最適なパラメーターの設定を発見し，そのパラメーターで学習し，分類器を作成し，テストデータを多値分類の手法を用いて分類し，その結果を評価する．

本実験では，同一のデータセットを線形カーネルと RBF カーネルを利用し，F タームによる特許自動分類における多値分類の問題に対して One-vs.-rest 法を用いて，F タームを自動で付与させ，それによって得られる精度を検証するものである．そして，先行研究において，Rikitoku や Li 等らにて採用されている $tf \times idf$ の重みを付けたベクトル空間モデルを採用することで，同一の質を持ったデータセットを利用することで，本実験の有効性を確認する．

5.2　データセット

今回実験に用いるデータは NTCIR が提供しているデータコレクションを使用する．NTCIR のデータコレクションについては，表 5.1 で詳細を解説する．NTCIR から提供されるデータコレクションの内，F タームに基づく特許自動分類に使われるデータは 1993 年から 1997 年のデータを訓練データとし，1998 年と 1999 年のデータ，総数 21,606 件をテストデータとしている．このデータは 108 個のテーマコードを持っており，1 つのテーマコードに付き，平均で 200 件ほどの特許がある．本実験では比較的新しい年代のデータを用いた実験を行なった．そのため，NTCIR で F タームに基づく特許自動分類に使われているデータではなく，NTCIR が提出されているデータの中から，別のデータを実験に用いた．本実験では，テーマコード「5B034」を選んだ．これは含まれる F タームは比較的少なく，分類においても実験，評価が容易に行えるため，このテーマコードを選択した．実験に用いたデータは表 5.2 に記載している．訓練データとして 1997 年から 1999 年を用

いて，テストデータには2000年のデータを用いる．訓練データの概要については，表5.3，テストデータの概要については表5.4を参照．

表 5.1: NTCIR データコレクション

ジャンル	年度	文書数	言語
特許全文	1993年～2002年	3,496,252	日本語
特許全文	1993年～2002年	3,496,252	英語
特許抄録	1993年～2002年	3,496,252	英語

表 5.2: テーマコード「5B034」のデータ

	件数
1997年	183
1998年	92
1999年	107
2000年	111

表5.3，表5.4は各Fタームごとに，そのFタームを持っている特許がどれくらい存在するかという「正の数」とそれぞれのデータ全体の割合を記載した表である．本実験では，学習データとして使われるFターム33個の内，各Fタームにおける「正の数」が33件以下しか存在しないデータは不均衡データとし，そのFタームでの特許自動分類の実験は行わない．結果，実験に使うFタームは15個とし，その15個のFタームにおいて特許自動分類の実験を行う．学習データとして選択された15個はテストデータにおいても同じFタームを用いる．表5.5が実験に用いるデータである．

表 5.3: テーマコー「5B034」の F タームの訓練データ

F タ ーム	正の数 (件) /全学習データ	割合
5B034AA01	92 /382	24%
5B034AA02	24 /382	6%
5B034AA04	68 /382	18%
5B034AA05	22 /382	6%
5B034BB00	1 /382	0%
5B034BB01	82 /382	21%
5B034BB02	69 /382	18%
5B034BB03	13 /382	3%
5B034BB04	10 /382	3%
5B034BB05	37 /382	10%
5B034BB06	3 /382	1%
5B034BB11	28 /382	7%
5B034BB13	1 /382	0%
5B034BB15	35 /382	9%
5B034BB16	13 /382	3%
5B034BB17	123 /382	32%
5B034CC00	16 /382	4%
5B034CC01	147 /382	38%
5B034CC02	48 /382	13%
5B034CC03	15 /382	4%
5B034CC04	28 /382	7%
5B034CC05	53 /382	14%
5B034CC06	34 /382	9%
5B034DD00	5 /382	1%
5B034DD01	120 /382	31%
5B034DD02	89 /382	23%
5B034DD03	28 /382	7%
5B034DD04	3 /382	1%
5B034DD05	98 /382	26%
5B034DD06	31 /382	8%
5B034DD07	75 /382	20%
5B034DD08	6 /382	2%
5B034DD09	1 /382	0%

表 5.4: テーマコー「5B034」の F タームのテストデータ

F ターム	正の数 (件) /全テストデータ	割合
5B034AA01	9 /111	8%
5B034AA02	11 /111	10%
5B034AA04	4 /111	4%
5B034AA05	6 /111	5%
5B034BB00	2 /111	2%
5B034BB01	5 /111	5%
5B034BB02	56 /111	50%
5B034BB03	8 /111	7%
5B034BB04	6 /111	5%
5B034BB05	6 /111	5%
5B034BB06	3 /111	3%
5B034BB11	7 /111	6%
5B034BB13	4 /111	4%
5B034BB15	8 /111	7%
5B034BB16	7 /111	6%
5B034BB17	15 /111	14%
5B034CC00	5 /111	5%
5B034CC01	89 /111	80%
5B034CC02	41 /111	37%
5B034CC03	41 /111	37%
5B034CC04	42 /111	38%
5B034CC05	56 /111	50%
5B034CC06	52 /111	47%
5B034DD00	49 /111	44%
5B034DD01	61 /111	55%
5B034DD02	72 /111	65%
5B034DD03	73 /111	66%
5B034DD04	75 /111	68%
5B034DD05	99 /111	89%
5B034DD06	106 /111	95%
5B034DD07	110 /111	99%
5B034DD08	111 /111	100%
5B034DD09	111 /111	100%

表 5.5: 実験に用いる F ターム

F ターム
5B034AA01
5B034AA04
5B034BB01
5B034BB02
5B034BB05
5B034BB15
5B034BB17
5B034CC01
5B034CC02
5B034CC05
5B034CC06
5B034DD01
5B034DD02
5B034DD05
5B034DD07

5.3 特許自動分類の評価方法

実験の評価は，本稿では先行研究や他の分類においても実験結果を評価する方法として用いられている適合率，再現率，F 値の計算を行うことによって実験結果の評価を行う．F 値とは予測結果の評価尺度の一つである．正負の 2 値分類の問題において SVM のような分類器の予測結果と真の結果を 5.6 の表のように表現する．

表 5.6: F 値

		真の結果	
		正	負
予測結果	正	TP	FP
	正	TP	FP

表 5.6 を用いて，適合率 (precision)，再現率 (recall)，F 値 (F-score) を解説していく．

- 適合率 (precision)：正と予測したデータの内，実際に正であるものの割合

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

- 再現率 (recall)：実際に正であるものの内，正であると予測されたものの割合

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

- F 値 (F-score)：適合率と再現率を総合的に評価する

$$\frac{2 \times Recall \times Precision}{Recall + Precision} \tag{5.3}$$

5.4 パラメーターのチューニング

本実験では，最適なパラメーターを得るために $K-$ 分割交差検定とグリッドサーチ法を用いて，実験に用いる 15 個の F タームそれぞれに最適なカーネル関数のパラメーターを導く．

5.4.1 $K-$ 分割交差検定 (K-fold cross-validation)

交差検定とは，与えられたデータを K 分割し，$K-1$ を学習データとし，残りをテストデータし，これを K 回検定を行い，得られた結果を平均することである 1 つの推定を得ることができる．この特徴から，SVM やカーネル手法におけるカーネル関数のパラメーターのチューニングにおいて最適なモデルを検出することに利用出来る [32]．本稿では，$K-$ 分割交差検定において，一般的に行われている $k = 10$ に設定し，パラメーターのチューニングを行い，SVM とカーネル関数の最適なモデルを設定する．

5.4.2 グリッドサーチ法

本実験では，線形カーネル，RBF カーネルのパラメーターのチューニングを行い，最適なパラメーターを得る．線形カーネルにおいては，最適なパラメーターのチューニングをヒューリスティックに行った．グリッドサーチ法という最適なパラメーターを網羅的に発見する方法があり，その手法を実装したオープンソフトウェアで，Python 言語で記述されているプログラムを実行することで RBF カーネルの最適なパラメーターを得ることができる [32]．本稿においても，この手法を利用した．この手法を利用する際は $K-$ 分割交差検定 ($K = 10$) も用いた．線形カーネルと RBF カーネルの最適なパラメーターは，5.7，表 5.8 で表している．

表 5.7: 線形カーネル:グリッドサーチ法，$K-$ 分割交差検定 (K = 10)

F タターム	Cost Parameter
5B034AA01	0.03125
5B034AA04	0.03125
5B034BB01	0.125
5B034BB02	0.03125
5B034BB05	0.03125
5B034BB15	0.5
5B034BB17	0.03125
5B034CC01	0.125
5B034CC02	0.03125
5B034CC05	0.125
5B034CC06	0.03125
5B034DD01	0.03125
5B034DD02	0.03125
5B034DD05	0.125
5B034DD07	0.03125

表 5.8: RBF カーネル:グリッドサーチ法，$K-$ 分割交差検定 (K = 10)

F タターム	C パラメーター	γ パラメーター
5B034AA01	0.5	0.0078125
5B034AA04	0.03125	0.0078125
5B034BB01	512	0.00012207
5B034BB02	0.03125	0.0078125
5B034BB15	2.0	0.5
5B034BB17	2.0	0.5
5B034CC01	2048	0.000030518
5B034CC02	2.0	0.0078125
5B034CC05	512	0.00012207
5B034CC06	0.5	0.0078125
5B034DD01	512	0.00012207
5B034DD02	0.03125	0.0078125
5B034DD05	8.0	0.0078125
5B034DD07	0.03125	0.0078125

5.5　実験の手順

テーマコード「5B034」のデータ抽出

実験に用いるテーマコード「5B034」を持ったデータセットを NTCIR のデータコレクションから抜き出し，与えられている F タームも同時に抜き出す (図 5.1).

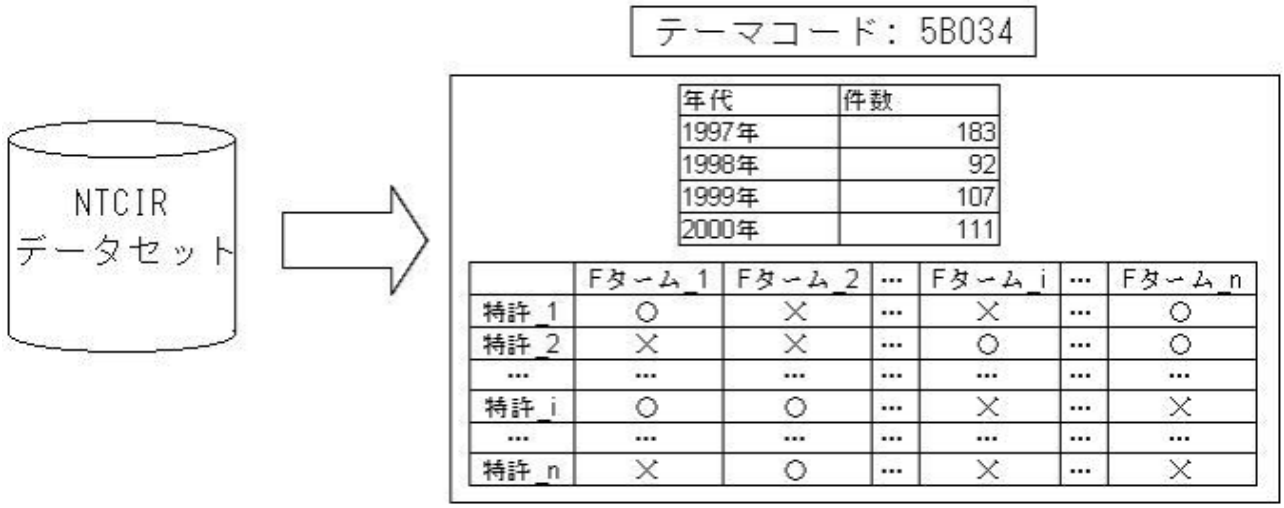

図 5.1: テーマコード「5B034」の抽出

$tf \times idf$ の計算

その変形させたデータセットからベクトル空間モデルを生成するために，最小単位の単語に分割する形態素解析のツールである MeCab 0.92 [1]を使用する．形態素解析に使用する項目は要旨と全請求項に限定する．その際に，分析された品詞が名詞のものだけを選び，また名詞が連続していた場合それらを連結し名詞句とすることで，単語としての意味情報を保持できるからである [21, 20]．また，名詞または名詞句の出現回数が 3 回に満たない場合，その名詞また名詞句は利用しない．また名詞の長さが 3 に満たない場合も同様である．このような条件の下で形態素解析を行った．その後，$tf \times idf$ を計算し，ベクトル空間モデルを生成した (図 5.2)

不均衡データの抽出

各 F タームの不均衡データを抽出し，実験に用いるデータのみに絞り込む.

[1]http://mecab.sourceforge.net /

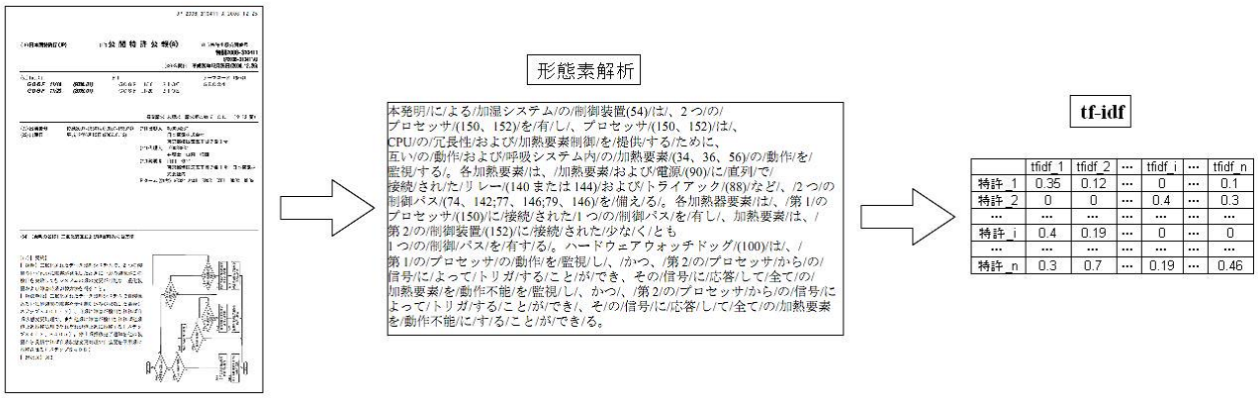

図 5.2: $tf \times idf$ の作成

パラメーター設定

グリッドサーチ法と $K-$ 分割交差検定 (K = 10) を用いて，最適なパラメーターの設定を探す.

One-vs.-rest 法を用いた実験

図 5.3 に示すような方法を用いて，各々の F タームごとに，最適なパラメーターを使用した線形カーネル，RBF カーネルを使い学習させ，テストデータに F タームのラベルを与える.

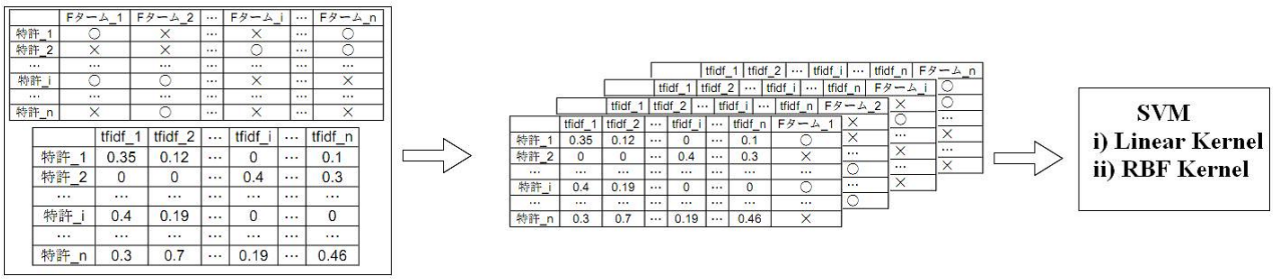

図 5.3: One-vs.-rest 法

評価

そして，得られた結果から，元々持っている F タームとを比較することで，F 値の値を得る．そして，その F 値によって本稿の実験の精度を評価する (図 5.4)

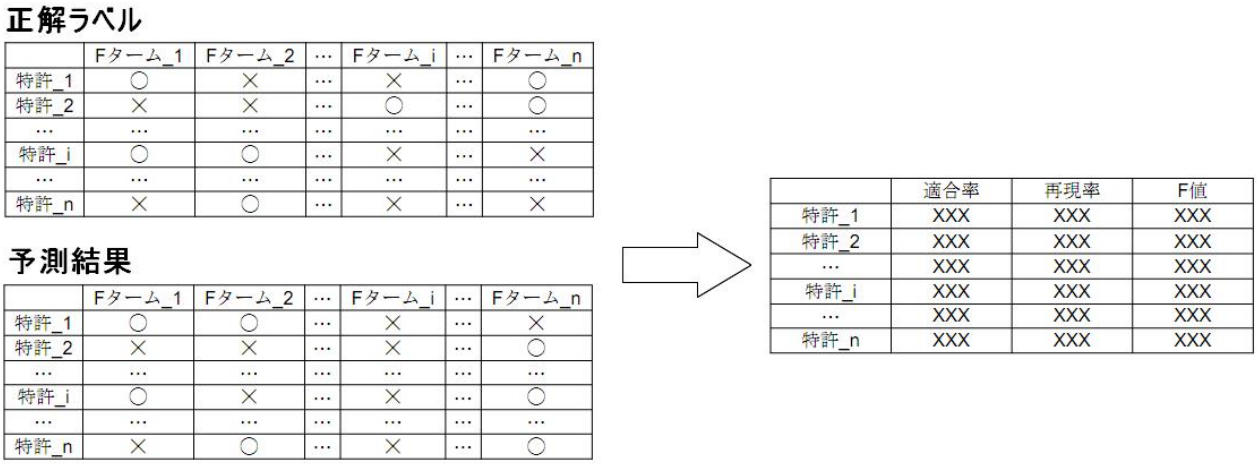

図 5.4: 評価方法

5.6 実験の構成

本稿において，データセットの抽出，データセットの生成，ベクトル空間モデルの生成を全て自分のコード (Python) によって作成し，そのデータを下に実験を遂行した．実験を行なった環境は Proccessor: 4 × 2.5GHz PowerPC G5，Memory: 1GB DDR2 SDRAM の下，実験を行なった．SVM を実装したソフトウェアは LIBSVM(version2.91)[2][32] を使用した．

5.7 実験結果

実験結果は線形カーネル，RBF カーネル，それぞれ表 5.9，表 5.10 で示す．表 5.9，表 5.10 では，いくつかの F タームでは分類が行われない実験が存在する．評価方法の節で解説した適合率，再現率，F 値の平均を表 5.11 に表示する．表 5.11 から RBF カーネルが線形カーネルより若干よい精度を上げていることがわかる．しかしながら，双方の実験結果において，分類が正常に行われない実験が多く，今回の実験からは線形カーネルと RBF カーネルにおける性能の違いを確認するには至らなかったが，限られた特許数の上，不均衡データが多くを占めるデータにおいてカーネル手法による実験の可能性を示すことができたが，カーネル手法を適応することで特許の自動分類の精度の向上を示せることをできななかった．

[2]http://www.csie.ntu.edu.tw /~ cjlin /libsvm /

表 5.9: 線形カーネルの結果

F ターム	TP	FP	FN	TN	適合率	再現率	F 値
5B034AA01	0	0	9	102	N /A	N /A	N /A
5B034AA04	0	0	4	107	N /A	N /A	N /A
5B034BB05	0	0	6	105	N /A	N /A	N /A
5B034CC02	0	0	41	70	N /A	N /A	N /A
5B034CC06	0	0	52	59	N /A	N /A	N /A
5B034DD02	0	0	72	39	N /A	N /A	N /A
5B034BB01	0	6	5	100	0.0	0.0	0
5B034BB02	0	2	56	53	0.0	0.0	0
5B034BB15	0	2	8	101	0.0	0.0	0
5B034DD01	0	4	61	46	0.0	0.0	0
5B034BB17	1	2	14	94	0.333333333333	0.0666666666667	0.111111111111
5B034CC01	17	3	72	19	0.85	0.191011235955	0.311926605505
5B034CC05	1	1	55	54	0.5	0.0178571428571	0.0344827586207
5B034DD05	1	0	98	12	1.0	0.010101010101	0.02
5B034DD07	2	0	108	1	1.0	0.0181818181818	0.0357142857143

5.8 議論

本実験では，不均衡データを除いたデータをそのまま実験に用い，データの正確性を考慮すると不均衡データを除いたデータを抽出し，$tf \times idf$ の計算を行い，実験する．このことについて，本実験では，時間的な都合上，このような処理が行われなかった．また，カーネル手法を最適なパラメーターで利用してさえも特許自動分類が行われてなかったデータについては，学習データとテストデータの特徴ベクトルに共通性が見られなかったことが考えられ，これは特許が前の技術や発明より新しい物を考えているという特徴を考えると，以前の特許と新たな特許とでは単語間においての共通項が少ないと考えられる．また，実験に使用された特許数が非常に少ないことが分類が正常に行われない結果を導いたと考えられることから，実験に使用するデータの数を増やしていくことが必要である．しかし，適切なデータ数はヒューリスティックに行う必要性が出てくると考えられ，その際に計算量が n^2 オーダーの計算が必要になってくることから実験にかかる時間は相当な時間がかかると予想され，時間が足りず実験数を増やして再実験を行うには至らなかった．

表 5.10: RBF カーネルの結果

F ターム	TP	FP	FN	TN	適合率	再現率	F 値
5B034AA01	0	0	9	102	N /A	N /A	N /A
5B034AA04	0	0	4	107	N /A	N /A	N /A
5B034BB02	0	0	56	55	N /A	N /A	N /A
5B034BB05	0	0	6	105	N /A	N /A	N /A
5B034BB15	0	0	8	103	N /A	N /A	N /A
5B034BB17	0	0	15	96	N /A	N /A	N /A
5B034CC02	0	0	41	70	N /A	N /A	N /A
5B034CC06	0	0	52	59	N /A	N /A	N /A
5B034DD02	0	0	72	39	N /A	N /A	N /A
5B034DD07	0	0	110	1	N /A	N /A	N /A
5B034BB01	0	6	5	100	0.0	0.0	0
5B034CC01	17	3	72	19	0.85	0.191011235955	0.311926605505
5B034CC05	1	1	55	54	0.5	0.0178571428571	0.0344827586207
5B034DD01	5	8	56	42	0.384615384615	0.0819672131148	0.135135135135
5B034DD05	18	4	81	8	0.818181818182	0.181818181818	0.297520661157

表 5.11: 適合率，再現率，F 値の平均

カーネル関数	適合率	再現率	F 値
線形カーネル	0.17	0.03	0.05
RBF カーネル	0.25	0.02	0.03

第6章　結論

本稿では，特許を形態素解析し，各文書における単語の重要度を計算する $tf \times idf$ 法を用いてデータセットを作成し，F タームによる特許自動分類において，カーネル手法を適用し，最適なパラメーターを設定することで，カーネル手法の特許自動分類における最適なカーネル手法を示すことが目的であった．本稿では，線形カーネルと RBF カーネルの 2 つのカーネル関数を選択し，それぞれ，最適なパラメーターを設定し，双方の実験結果の比較を行うことを提案した．結果は，RBF カーネルでの実験から得られた結果と線形カーネルでの実験から得られた結果から RBF カーネルの実験が線形カーネルを利用した時より，精度の良い結果が得られた．線形カーネルの精度は RBF カーネルよりも低いが，特許自動分類にカーネル手法を適用することには有効性があることが示す結果となった．しかしながら，全ての実験において，特許自動分類が上手く機能しなかった．この研究を通して，特許自動分類における問題点を認識することができ，また，カーネル手法の有効性についても，最適なパラメーターを設定することを条件とすることで，カーネル手法の有効性を最大限に活かせることができる可能性があることを示すことができた．

今後の課題

この結果を踏まえ，今後の課題は，カーネル関数の選択，もしくは作成が必要であり，線形カーネルと RBF カーネル以外のカーネルを選択し，それぞれのパラメータを設定することで，よりよい精度になると考えられる．本稿における実験では，最適なパラメータを設定し，その結果，線形カーネル，RBF カーネルの双方において，精度を得られない結果に終わったということことから，カーネル関数の選択だけではなく，特許情報からベクトル空間モデル化する際の他の手法を提案することが必要になると考える．また，不均衡データをもつ F タームを除いて実験を行うことは，その F タームが自動で付与されないことになるため，本来の目的とは少々違う面もあるため，不均衡データにおいても通常の実験のようにできるようなアルゴリズムの提案や，工夫を考える必要がある．最適なパラメーターを検出しながら，全ての実験が上手く機能しなかった点を考慮すると，特許のテキスト情報に対して，潜在意味インデキシング (latent semantic indexing)，スペクトルカーネル，または，特許が構造化された情報であることから，構造化に対するカーネル，例えば木カーネルの適用を考慮する必要がある．また，本稿における実験では，実験に用いる特許のデータ数が少なくそのほとんどが不均衡データという状態での実験になったことから，今後は不均衡データを除き，また，データ数もある程度確保した上で同様の実験

を行うことを提案する．また，不均衡データを含んだ状態で，適切なカーネル手法のアルゴリズムを考慮することも提案する．最後に，SVM，k-nn やカーネル手法など様々な手法の改善と向上を行うことで特許の自動分類の精度の向上を図ることは機械学習全般の研究の発展につながると考えられ，また機械学習の研究の成果を特許の自動分類に適応することで分類精度の向上が図られると考えられることから，双方の成果を共有することを提案したい．

謝辞

本稿を進めるにあたり，様々なご指導を頂きました主指導教員である Ho Tu Bao 教授に感謝致します。また、日常の議論を通じて多くのご助言とご支援を頂いた河崎さおり助教をはじめ，多くの知識や示唆を頂いた研究室の皆様に感謝します。

参考文献

[1] 知的財産戦略会議, 知的財産戦略大綱知的財産戦略会議, 2002

[2] 間瀬久雄, 文書内の言語構造を利用した特許文書分類・検索技術の研究, 名古屋大学博士学位論文, 2007.

[3] 日本国特許庁, 特許行政年次報告書 2009 年版特許庁, 2009

[4] 日本国特許庁, 特許審査迅速化の実施計画 特許審査迅速化・効率化のための平成 21 年度の取組について http://www.jpo.go.jp/cgi/link.cgi?url=/torikumi/zinsoku/h21zinsoku_plan.htm May. 17, 2010

[5] M. Iwayama, A. Fujii, and N. Kando (2005), Overview of classification subtask at NTCIR-5 patent retrieval task, In Proceedings of NTCIR-5 Workshop Meeting.

[6] M. Iwayama, A. Fujii, and N. Kando (2007), Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task, In Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access (NTCIR '07).

[7] M. Rikitoku (2007), F-term classification Experiments at NTCIR-6 for Justsytems Proceedings of the 6th NTCIR Workshop Meeting, 2007. ACM Trans. Asian Lang. Inform. Process., Vol. 7, No. 2.

[8] Y. Li, K. Bontcheva, and H. Cunningham (2007), SVM Based Learning System for F-term Patent Classification Proceedings of the 6th NTCIR Workshop Meeting, 2007.

[9] Y. Li, and K. Bontcheva (2008), Adapting Support Vector Machines for F-term-based Classification of Patents ACM Transactions on Asian Language Information Processing, Vol. 7, No. 2, Article 7, June. 2008.

[10] N. Cristianini, and J. Shawe-Taylor, 大北剛訳 (2006), サポートベクターマシン入門, 共立出版

[11] J. Shawe-Taylor, and N. Cristianni (2004), Kernel Methods for Pattern Analysis, Cambridge university Press.

[12] J. Shawe-Taylor, and N. Cristianni 著, 大北剛訳, カーネル法によるパターン解析, 共立出版

[13] L. Zhang, D.Zhang, S. J. Simoff, and J. Debenham (2006), Weighted Kernel Model for Text Categorization Fifth Australasian Data Mining Conference, 2006.

[14] F. Colas,, P. Paclik, J. Kok, and P. Brazdil (2007), Does SVM Really Scale Up to Large Bag of Words Feature Spaces? Lecture Notes in Computer Science, 2007, Volume 4723, pp296-307, 2007.

[15] N. Cancedda, N. Cesa-Bianchi, A. Conconi, G. Claudio, C. Goutte, Y. Li, J. M. Renders, J. Shawe-Taylor, and A. Vinokourov (2002), Kernel Methods for Document Filtering The Eleventh Text Retrieval Conference, 2002.

[16] M. Murata, T. Kanamura, T. Shirado, and H. Isaharam (2007), Using the K-Nearest Neighbor Method and SMART Weighting in the Patent Document Categorization Subtask at NTCIR-6 Proceedings of the 6th NTCIR Workshop Meeting, 2007.

[17] 日本特許庁 (2010), 平成 22 年度知的財産権制度説明会 (初心者向け) テキスト日本国特許庁, 2010

[18] 独立行政法人　工業所有権情報・研修館　「パテントマップガイダンス　検索項目の概要」, http://www.ipdl.inpit.go.jp/HELP/pmgs/database/format_summary.html 最終アクセス 2011 年 2 月 4 日

[19] T. Joachims (1998), Text Categoraization with Support Vector Machines: Learning with Many Relevant Features European Conference on Machine Learning, pp.137-142.

[20] 内山清子 (2009), 特許文における複合語の扱いについて, 一般財団法人日本特許情報機構, Japio 2009 YEARBOOK.

[21] 内山清子 (2007), 特許文書における複合語の意味関係解析, 一般財団法人日本特許情報機構, Japio 2007 YEARBOOK.

[22] 間瀬久雄, 辻洋, 絹川博之, 石原正博 (1998), 特許テーマ分類方式の提案とその評価実験, 情報処理学会, Vol. 39, No. 7, pp2207–2216.

[23] 日本国特許庁. 2008, 国際特許分類，FI，F タームの概要とそららを用いた先行技術調査 日本国特許庁.

[24] 日本国特許庁 (2009) 国際特許分類 (2009 年バージョン)　指針 日本国特許庁.

[25] H. Tanabe (2008), Prediciting Protein-Protein Interactions Using Multiple Kernel Learning Japan Advanced Institute of Science and Technology, 2008

[26] 赤穂 昭太郎 (2008), カーネル多変量解析—非線形データ解析の新しい展開, 岩波書店

[27] 中野 桂吾 (2004), 日本語固有表現抽出における文節情報の利用 筑波大学大学院博士課程システム情報工学研究科修士論文, 2004

[28] G. Richter, and A. MacFarlane (2005), The impact of metadata on the accuracy of automated patent classification World Patetn Information, Vol. 27.

[29] I. Schellner (2002) Japanese File Index classification and F-terms World Patent Information, Vol. 24, pp.197-201

[30] J. Rousu (2006), Kernel-Based Learning of Hierarchical Multilabel Classification Models Journal of Machine Learning Research, Vol. 7, pp.1601-1626K.

[31] K, Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholokopf (2001), An Introduction to Kernel-Based Learning Algorithms IEEE Transactions on Neural Networks, Vol. 12.

[32] C-W Hsu, C-C Chang, and C-J Lin (2003), A practical guide to support vector classification, Department of Computer Science, National Taiwan University.