# 第 1 章

# はじめに

## 1.1 研究の背景

データベース・システムにおける情報検索，および，知識発見のための主要な基本操作は連想検索である．ここで，連想検索とは，あるキーワードに関連する情報をそのキーワードが表すアドレスではなく，そのキーワードの内容に応じて検索することをいう [19, 26]．現行のデータベース・システムにおける連想検索は，パターン・マッチングによる検索であり，異なる表現形態であるが同一の意味をもつデータや近い意味をもつデータの検索を行うことはできない [18]．また，同一のデータがもつ多義性を取り扱うことはできない．データ間の意味的な関係の扱いについては，データ間の関係を静的かつ明示的に記述し，同一性，相異性を判定する方法が広く用いられてきた [5, 9]．しかし，その判定は，静的に与えられた関係を用いて，曖昧性を含んで行われる．例えば，シソーラスを用いて同義語を照会する方法があるが，その同義語は，シソーラスの設計時に静的に決定され，また，同義であることの定義には曖昧性を含んでいる．また，各単語について，その意味を表すベクトルを形成し，ベクトル間の相関の強さによって意味の近さを判定する方式がある [11]．その方式においては，ある単語に意味的に近い単語を求めるために，他の全単語各々に対応するベクトルとの間で相関の強さを計算しなければならない．

我々は，データ間の意味的な同一性，相異性は，静的な関係によって決定されるのではなく，文脈や状況に応じて動的に変化するものであり，その動的な要素を含んで決定しなければ，データ間の関係の曖昧性を排除することはできないいものと考える．このようなデータ間の意味的な関係を文脈に応じて動的に計算するモデルとして，意味の数学モデルが提案されている [12, 15, 16]．

意味の数学モデルは，ほぼ無限通りの文脈や状況に応じて動的に変化するデータ間の意味的な関係を計算することを目的としたモデルである．このモデルに基づいた連想検索により，文脈に応じて，検索キーワードに意味的に近い検索対象語を連想検索することが可能となる．検索対象となるデータが大量となった場合には，意味の数学モデルによる意味的連想検索の高速化が必要になる．意味の数学モデルは，ある単語に意味的に近い単語を，大量のデータの中から高速に抽出する能力を潜在的に備えている．これは，このモデルにおいて，意味の近い 2 つの単語は，空間上における距離の近い 2 点として配置されていることによる．

本論文では，意味の数学モデルによる意味的連想検索において，文脈に応じた高速な連想検索を行なうためのアルゴリズムを提案し，その実現方式を示す．

パターン認識，計算幾何学の分野においては，比較対象あるいは検索対象のデータ集合の各要素が固定的な位置にあることを利用してボロノイ図を予め計算し，ボロノイ図を利用して高速化を実現する方法が提案されている [42, 43]．また，空間データベース (spatial database) の分野においては，検索対象となるデータ（オブジェクト）が固定的な位置に存在することを利用し，それらのデータのグルーピング（クラスタリング）を行い，インデックスを作成することにより，検索の高速化を行っている [10, 7]．

これらの高速化方式との比較において，本連想検索方式の高速化の実現では，動的に変化するデータ間の位置関係を前提としたアルゴリズムの設計が必要となる．本連想検索方式では，正規直交空間上に，検索者が発行する検索キーワード（検索語ベクトル）および文脈語に対応するベクトル，および，検索対象（比較対象）データに対応するベクトル（比較対象語ベクトル）を写像する．

そして，検索キーワード，および，その文脈を確定するための文脈語群が与えられると，その文脈に対応する正規直交空間の部分空間を動的に選択する．その部分空間における検索語ベクトルと比較対象語ベクトルの位置により，その部分空間において検索キーワードと意味的に最も近い比較対象データを動的に抽出する．このように，本連想検索方式では，文脈理解を正規直交空間における部分空間選択によって実現し，与えられた文脈，すなわち，選ばれた部分空間において検索キーワードに最も近い比較対象データを抽出する．

この連想検索機構では，文脈に対応する部分空間上において，検索キーワードと比較対象語の位置的な近さを計算するので，各検索毎に，対応するベクトル群を，選択された部分空間上に動的に写像する．すなわち，検索毎に動的にベクトル群の位置関係が

決まるので，検索の高速化を実現するために静的なインデックスを予め作成しておくことはできない．そこで，本連想検索においては，文脈に応じて動的に変化するデータ間の関係を前提としたアルゴリズムを実現する．

本方式は，多変量解析による空間生成を用いた情報検索手法 [27] とは，次の点で本質的に異なる．本方式では，直交空間における部分空間の選択を行う演算（意味射影）を定義し，その演算により，言葉の意味を文脈に応じて，曖昧性を排除して解釈する機構を実現している．この機構により，検索キーワードと検索対象データの間の意味的な関係を，与えられた文脈に応じて動的に計算することを可能としている．

また，近年，データベースとネットワークが普及し，多くの人は，様々なデータベースの利用が可能になっている．しかし，多数のデータベース群は，異なるデータベース設計者により構築されている．そのため，データベースに対する思想は元より，同じ意味を持つデータが，異なる表現を持つことがある．一般に利用されているパターン・マッチングによる連想検索，すなわち，キーワードを入力し，そのキーワードと同じパターンを持つデータを抽出する機構においては，異なる表現を持ち，かつ，同じ意味を持つデータを判別できないので，そのようなデータを対象とする検索を行うことは困難である．特に，様々なデータベースを統一的に扱うことを目的としたマルチデータベース・システムの実現においては，データの持つ意味を扱うことが可能な検索方式が必要となる．本論文では，パターン・マッチングではなく，意味的な等価性，類似性に関する計算によりデータベース検索を行う意味的連想検索方式の高速化アルゴリズムを提案する．本アルゴリズムは，意味的連想検索方式として，意味の数学モデルを適用した方式を対象とする．意味の数学モデルによる意味的連想検索の特徴は，文脈や状況に応じた意味の扱いを実現する点にある．本アルゴリズムは，キーワード，および，それを説明する文脈語列を受けとることによって意味的に関連する情報を高速に抽出するために用いられる．本論文では，意味的連想検索を実現するために，意味の数学モデルを適用した高速検索方式について検討を行った．意味の数学モデルによる連想検索の特徴は，文脈や状況に応じた意味の扱いが可能なことである．

本論文では，意味的連想検索の高速なアルゴリズムを提案し，意味の数学モデルによる高速な連想検索システムを実現した．また，高速な意味的連想検索のアルゴリズムの有効性を確認するために，英英辞書における基本英単語を対象とした実験を行い，高速な意味的連想検索のアルゴリズムが有効であることを確認した．また，実在するデータベース群を対象とした，意味の数学モデルによるマルチデータベース・システムへの

適用実験を行った．マルチデータベース環境とは，様々なデータベースシステムを利用できる環境のことである．しかし，各々のデータベースシステムは，別々の環境で作成され，ポリシーも異なる．この様な環境では，利用者は，各データベースの操作方法を個別に習得せねばならず，また，探したいデータのキーワードもデータベースごとに異なるので，様々なキーワードを試して，各々のデータベースからデータを集めなくてはならない．

このような状況を改善するためには，

(1) 全てのデータベースを統一した環境に再構築する．

(2) 利用者にデータベースが統一されているようにみせる．

の方法が考えられる．

しかし，(1) の方法には，何十年もかけて構築された世界中のデータベースを再構築せねばならないといった欠点と共に，最終的なデータベース群を，どのように再構築するかを，誰かが決定せねばならないという欠点がある．(2) の方法は，今までのデータベースシステムには一切手を加えず，マルチデータベースシステム内で，データベースシステムごとの違いを吸収し，あたかも，一つのデータベースに見せる方法である．この場合，世界中のデータベースに手を加えることなく，利用者に，データベース群が統合された環境を提供することが可能である．本論文では，マルチデータベースシステムの一部に意味の数学モデルを適用し，さらに，高速検索アルゴリズムを適用した実験システムを作成した．その実験結果により，意味の数学モデルを適用したマルチデータベースシステムにおいても，文脈に応じた意味による検索が可能になることを確認し，高速な意味的連想検索のアルゴリズムが有効であることを確認した．さらに，本論文では，マルチメディアデータを対象としたリアルタイム意味的連想検索アルゴリズムについて述べる．このアルゴリズムでは，優先度の高いデータから計算を行い，指定された時間で検索を中断することで実時間性を実現する．さらに，このアルゴリズムを並列処理に対応させた，マルチメディアデータ検索並列アルゴリズムについて述べ，実際の画像から得たメタデータを対象とした実験を行った．実験結果より，本アルゴリズムの有効性を検証する．

## 1.2 研究の目的

意味の数学モデルは，ほぼ無限通りの文脈や状況に応じた意味的な関係を動的に計算することを目的としたモデルである．このモデルに基づいた連想検索システムにより，検索キーワードに意味的に近い検索対象語を連想検索することが可能となる．検索対象となるデータが大量となった場合には，意味の数学モデルによる意味的連想検索の高速化が必要になるが，単純なパターン・マッチングによる連想検索などで利用されているR-tree[7, 10] などの高速化を，動的に関係を計算するモデルにそのまま適用することはできない．また，ニューラルネットワークを使用した方法では，学習に必要な時間が無視できなくなる [7]．しかし，意味の数学モデルは，ある単語に意味的に近い単語を，大量のデータの中から高速に抽出する能力を潜在的に備えている [33]．意味の数学モデルにおいて，意味の近い 2 つの単語は，空間上における距離の近い 2 点として表されている．本論文は，この性質を利用して，意味の数学モデルによる意味的連想検索の高速化を実現する方式について検討を行った．

また，近年マルチメディアデータを対象とした，データベース作成の試みが行われている．しかし，従来のパターンマッチを基本としたデータベースでは，直接的に「赤色」などと指定するしか方法がなく，赤に近い橙色は検索結果として現れないといったことがおきる．そこで意味の数学モデルを利用して，「暖かい色」など，データの雰囲気を指定することにより，検索者は多くの情報を得ることができるのである．本論文では，マルチメディアデータを対象とした，実時間並列検索アルゴリズムに付いて述べ，実験により，プロセッサを増加が検索精度の向上に結び付くことを確認する．

## 1.3 本論文の構成

以下，2 章では，意味の数学モデルの基本モデルについて述べる．3 章では，意味の数学モデルによる高速な意味的連想検索のアルゴリズムを述べる．4 章では，意味の数学モデルを適用したマルチデータベース・システムについて述べ，高速意味的連想検索の有効性の評価に関する実験と，その結果について述べる．5 章では，意味の数学モデルによる実時間並列マルチメディアデータ検索アルゴリズムについて述べ，実験によってその有効性を確認する．6 章では，意味の数学モデルの基本モデルのシステム設計と実装について述べる．7 章において，結論を述べる．