１２－開－０２

データベース構築促進及び技術開発に関する報告書

ＧＩＳと全文検索エンジンを連携した文章管理システム

平成１３年３月

財団法人　　データベース振興センター

委 託 先　　株 式 会 社　　創 建

KEIRIN

この事業は、競輪の補助金を受けて実施したものである。

序

データベースは、わが国の情報化の進展上、重要な役割を果たすものと期待されている。今後、データベースの普及により、わが国において健全な高度情報化社会の形成が期待される。さらに海外に対して提供可能なデータベースの整備は、国際的な情報化への貢献および自由な情報流通の確保の観点からも必要である。現在わが国で流通しているデータベースの中でわが国独自のものは約半数であるが、わが国データベースサービスひいてはバランスある情報産業の健全な発展を図るためには、今後もわが国独自のデータベースの構築およびデータベース関連技術の研究開発を強力に促進し、データベースの拡充を図る必要がある。

このような要請に応えるため、(財)データベース振興センターでは日本自転車振興会から機械工業振興資金の交付を受けて、データベースの構築および技術開発について民間企業、団体等に対して委託事業を実施している。委託事業の内容は、社会的、経済的、国際的に重要で、また地域および産業の発展の促進に寄与すると考えられているデータベースの構築とデータベース作成の効率化、流通の促進、利用の円滑化・容易化などに関係したソフトウェア技術・ハードウェア技術である。

本事業の推進に当って、当財団に学識経験者の方々で構成されるデータベース構築・技術開発促進委員会（委員長　東海大学教授　上條史彦氏）を設置している。

この「ＧＩＳと全文検索システムを連携した文章管理システム」は、平成１２年度のデータベースの構築促進および技術開発促進事業として実施した課題の一つで、当財団が株式会社　創建に対して委託実施したものである。この成果が、データベースに興味をお持ちの方々や諸分野の皆様方のお役に立てば幸いである。

なお、平成１２年度データベースの構築促進および技術開発促進事業で実施した課題は次表のとおりである。

平成１３年　３　月

財団法人　データベース振興センター

平成１２年度　データベース構築・技術開発促進事業委託課題一覧

分　野	No.	課　題　名	企　業　名
一　般	1	学びに活用する簡易ＰＣサーバ・システム、マルチメディア・データベース	（株）トライアード・プロジェクト
	2	ＧＩＳと全文検索エンジンを連携した文章管理システム	（株）創建
	3	情報可視化によるドキュメント構造化の調査研究	（株）日本総合研究所
	4	先端産業の企業検索用シソーラス作成と企業検索システム構築	（株）日経リサーチ
	5	インターネット環境における博物館型地図画像データベースの構築	（財）地図情報センター
地域振興	6	地域住宅地図情報システム〔地域住宅ＧＩＳ〕	石井測量設計（株）
	7	阿蘇の楽しみデジタル図鑑作成	（財）阿蘇町地域振興公社
	8	Ｂ to Ｓ共働情報マッチングＤＢパイロットシステムの作成	（株）八幡コンピュータセンター
	9	Ｗｅｂ型ＧＩＳを利用したバリアフリーデータベースのプロトタイプ作成	（株）札幌ネクシス

目　次

1 はじめに

1．1 文書管理と地図について

近年、全文検索に対する需要が増加しており、各メーカー毎に活発な開発競争がなされている。今後、全文検索はインターネットだけの利用にとどまらず、組織で作成される文書の活用のためにも不可欠な技術として定着していくことになると考える。

現在、組織内で作成された文章の多くは、電子ファイルの形で文書として保存されている。本報告書およびシステムでは、この作成された内容を文章とし、保存されたファイルを文書と表現することとする。

組織内で作成される文書は、地名等の位置に関する情報を含んでいるものが多く見られる。これらを地図上に配置することができれば、対象となる地域の特徴が認識しやすくなる。

地図上にデータを配置する手法として、GIS（地理情報システム）がある。GIS は、地図を媒介として多様な情報を扱うことを可能にしている。

しかし、GIS でデータを扱うには位置を特定する情報が必要である。DB（データベース）の項目のように整理され、位置に関係する項目が特定できる場合は位置の特定は容易である。しかし、書式の決まっていない文書等は、位置に関する情報を内容を確認しながら、指定しなければならない。このような文書を GIS 上で扱うには多くの手作業が必要とされる。

そこで、これらの作業を全文検索により内容の検索を行い、地図との関連付け作業を自動的に行うシステムの作成を行うものである。

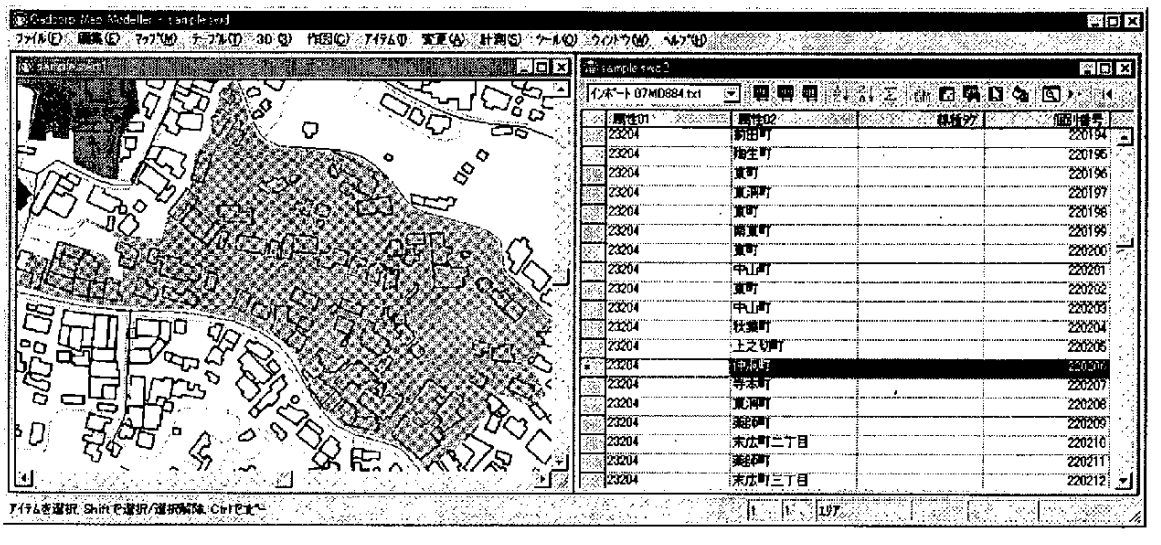

図１－１　位置の情報と関連する図形データ

1．2 システムの整備効果

本システムは、GIS と全文検索システムを連携させることにより、情報を整理することなく抽出し、地図上に表現するものである。本システムにより以下の効果が得られる。

①文書の多くを占める非定型情報と地理情報を同時に扱うことが可能になる。これにより柔軟に情報の関連性を見出すことができる。
②地図上からの検索と、文書情報からの検索どちらも可能になる。文書を探す場合に、登録されたキーワードの制約を受けることがなくなる。
③データベースの更新作業から開放される。文書の更新や登録を自動的に行うため、インデックス作成や、位置との関連付け作業が不要となる。
④全文検索が不得手としていた表現能力が、GIS により補える。従来の全文検索では、検索結果として文字情報しか返すことができなかったが、地図上に結果を展開することで視認性が向上する。

GIS と全文検索システムの連携によりそれぞれの特徴を最大限に生かすことが可能になる。不足する機能を互いに補完するシステムとなる。応用範囲として以下のような用途を想定している。

①計画策定支援システム

住民から行政に寄せられる意見等の情報は、場所に依存する場合が多い。これらの情報を計画策定時に活用することは重要な要素である。これらの意見を地域毎に集めるには、多くの作業が必要とされる。また、隣接する地域を扱うにはさらに多くの作業を行わねばならない。

そこで、本システムを用いて対象地域周辺の情報を集め、基礎資料として活用することが考えられる。また、業務を引き継ぐ場合にも蓄積されたデータがそのまま利用できるため、資料作成の負担が軽減される。

②営業支援システム

民間企業を対象とした営業支援システムである。営業履歴のほとんどは場所との関連を持ち、地図の利用頻度も高い。営業情報を地図に展開することで、個人の記憶に依存していた部分を共有することができ、機会や情報のロスを減らせるようになる。

1．3 システムの全体構成

本システムの全体構成は次のようになる。作成した文書は、GIS と全文検索システムにより利用される。将来的には、紙で管理されている情報も取り込み、イントラネットでどこからでも利用可能なシステムとする。

今回の開発は、破線の範囲内を対象とし、クライアント／サーバーで構成されるプロトタイプの作成を行う。

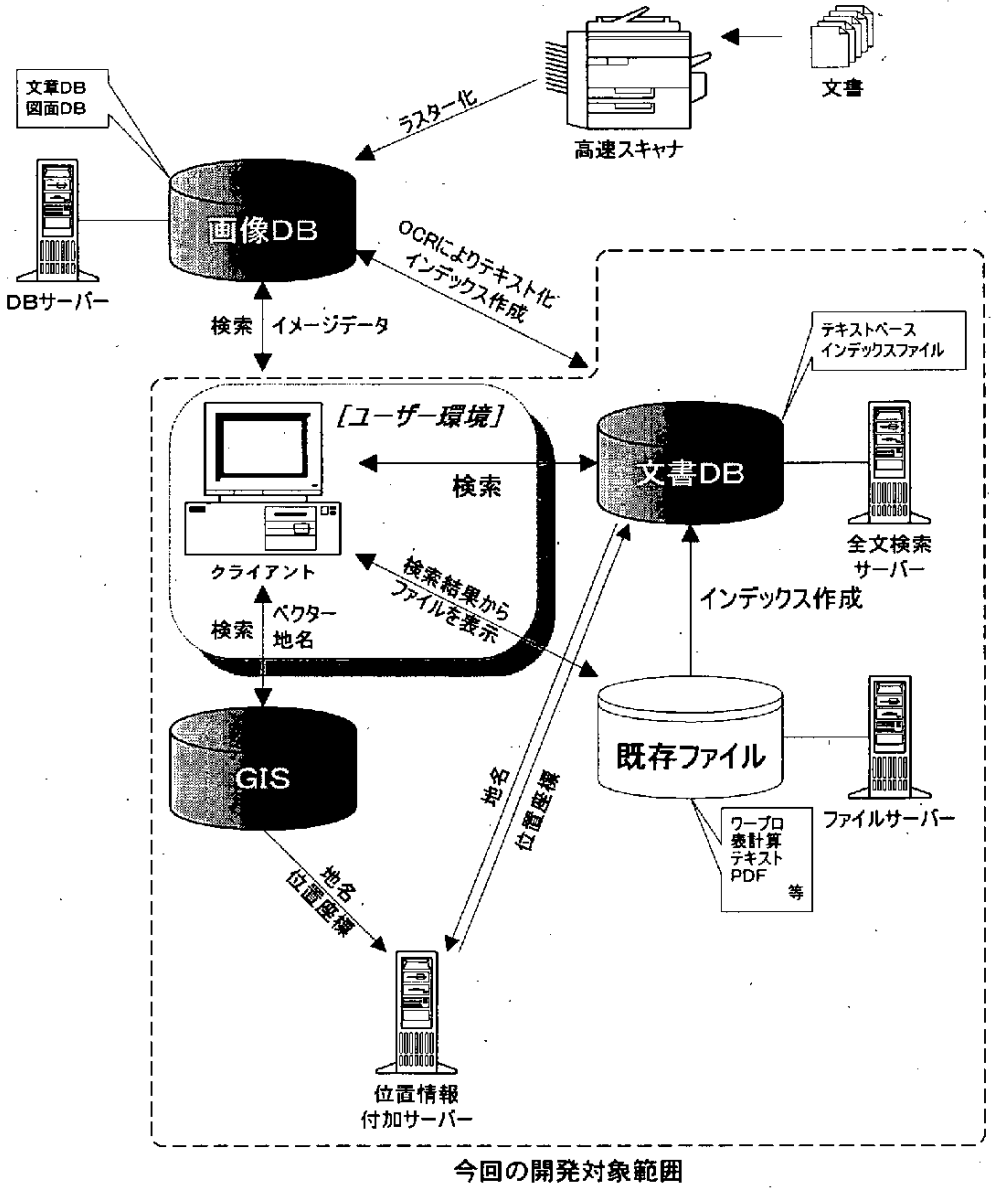

図１－２　システム全体構成

2 開発方針と開発手法

2．1 システムの開発方針

システムの開発にあたり、設計方針と開発方針を次のように定める。

（1） 設計方針

①既存技術の活用

今回は技術開発的な要素が強いシステムであるが、期間が限られている。そこで、既存技術を用いることで工期の短縮を図るとともに、汎用性が高く別の分野で応用可能なシステム設計を行うものとする。

②DB 項目の変更

今回の DB の構造は実験として様々なタイプを試す必要があり、DB 項目の変更が頻繁に行われる。そこで、DB 項目については大枠を決めるのみで随時必要に応じて変更を認めることとした。

（2） 開発方針

本システムは、プロトタイプ開発であるため、柔軟な開発体制をとることとした。

①初期設定ファイルの作成

ファイル構成、変数等を定義した初期設定ファイルは、直接書き換えることで対応する。メニューからの設定の変更は行わない。

②プロトタイプモデル

処理によってモジュールの分割を行い、プロトタイピング手法により、作成する。機能を確認しながら、段階的な機能の見直しを行う。

2．2 システム構成

開発にあたってのシステム概念と構成についてまとめる。

2．2．1 システム概念

システムに用いる技術は、GIS と全文検索技術である。ここでの全体構成は次のようになる。

作成した文書はファイルサーバーに格納された後、検索のためにインデックスが作成される。このインデックスは全文検索に用いるデータである。さらに、文書に対する位置座標を位置情報付加サーバーを通じて取得する。この時、位置情報は GIS データから読み込んだデータを用いる。

ユーザー環境からは蓄積された文書 DB に、サーバー経由でアクセスして関連する情報の検索結果を受け取る。地名や図形データは GIS データから取得し、文書情報は文書 DB から取得する。検索結果を受け、該当する文書をファイルサーバーより取り出す。

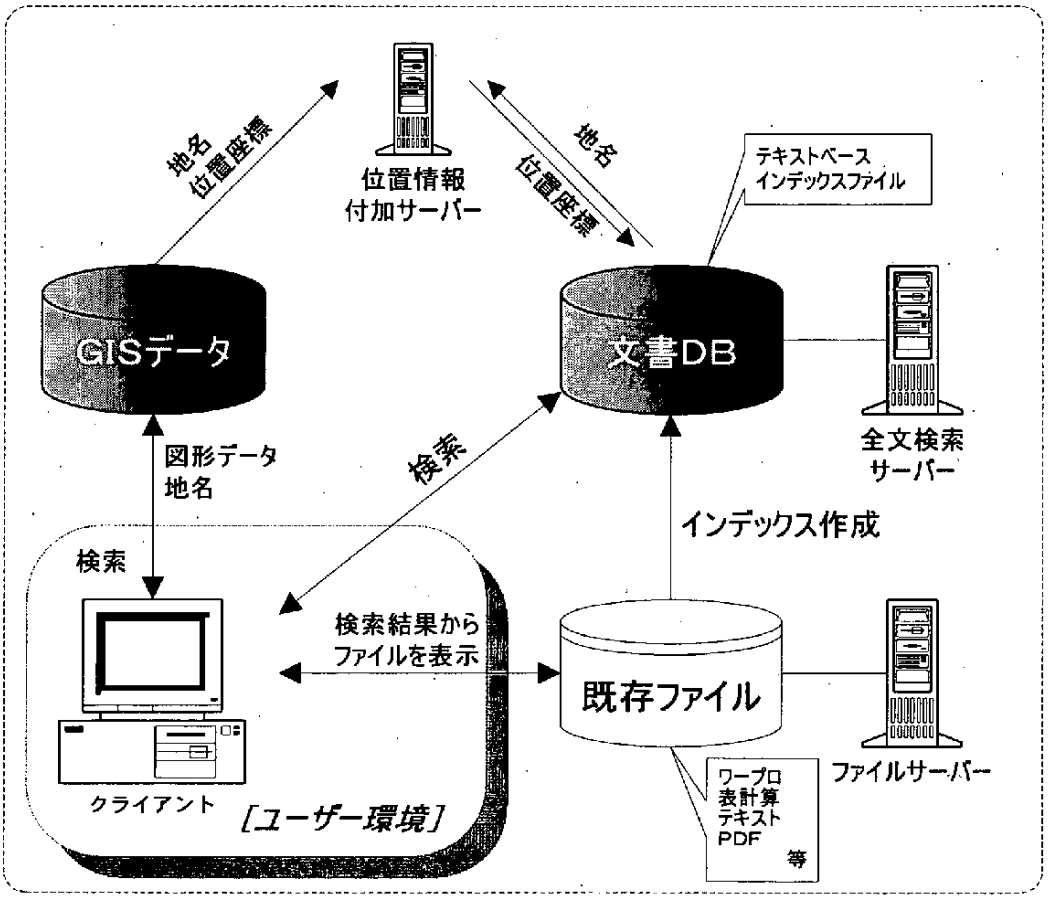

図２－１　システム構成図

（1） サーバーの機能

サーバーは概念的に以下の３種類に区分される。

①ファイルサーバー

ユーザーが作成したデータを保存する。ファイルサーバーそのものの機能はデータの保存のみで一般的なファイルサーバーと同様なものである。

②全文検索サーバー

ファイルサーバーに蓄積された文書からインデックスの作成を行う。インデックスは検索時に用いられるDBとして全文検索サーバー上に保存される。サーバー上にあるのはインデックスのみで文書データはファイルサーバー上にある。

インデックスの作成を行った後、位置情報付加サーバーに文書を渡し位置情報の問い合わせを行う。問い合わせ結果を受けて、インデックスに位置情報の追加を行う。

インデックスを作成後、全文検索サーバーは待ち状態に入り、クライアントからの検索要求を待つ。検索要求が発行されると、検索を実行し、該当する文書のファイル情報と位置情報を返す。

③位置情報付加サーバー

全文検索サーバーから、位置情報の問い合わせを受ける。全文検索サーバーから検索要求を受け取ると、地名情報を取出し、該当する位置情報を返す。位置情報付加サーバーはGISデータの地名と位置情報を利用する。このサーバーは、GIS データと全文検索サーバーのデータを利用し、サーバー固有のデータは持たない。

（2） クライアントの機能

クライアントは次の機能を持つ。

①検索条件の入力

検索条件の入力画面を表示し、条件の入力待ちを行う。入力された条件は処理フローに基づき検索が行われる。

②全文検索サーバーへの問い合わせ

ユーザーが入力した条件をもとに、サーバーが理解できる問い合わせ文を作成し、サーバーに問い合わせを行う。

③検索結果の表示

サーバーの回答を受けて、画面上に結果を表示する。

２．２．２ 開発手法

開発にあたり、インターフェース毎に機能を分割し、システムの開発を行う。

（１）処理フロー

本システムの開発要素は、ユーザー操作部、検索部、結果表示部に分けられる。さらに、検索部では、位置条件を扱う部分と文字条件から検索する部分とに分けられる。また、結果表示部においても、検索結果を地図上に展開する部分と、該当文書の表示を行う部分とに分けられる。

処理によってモジュールの分割を行い、モジュール単位で開発を行う。

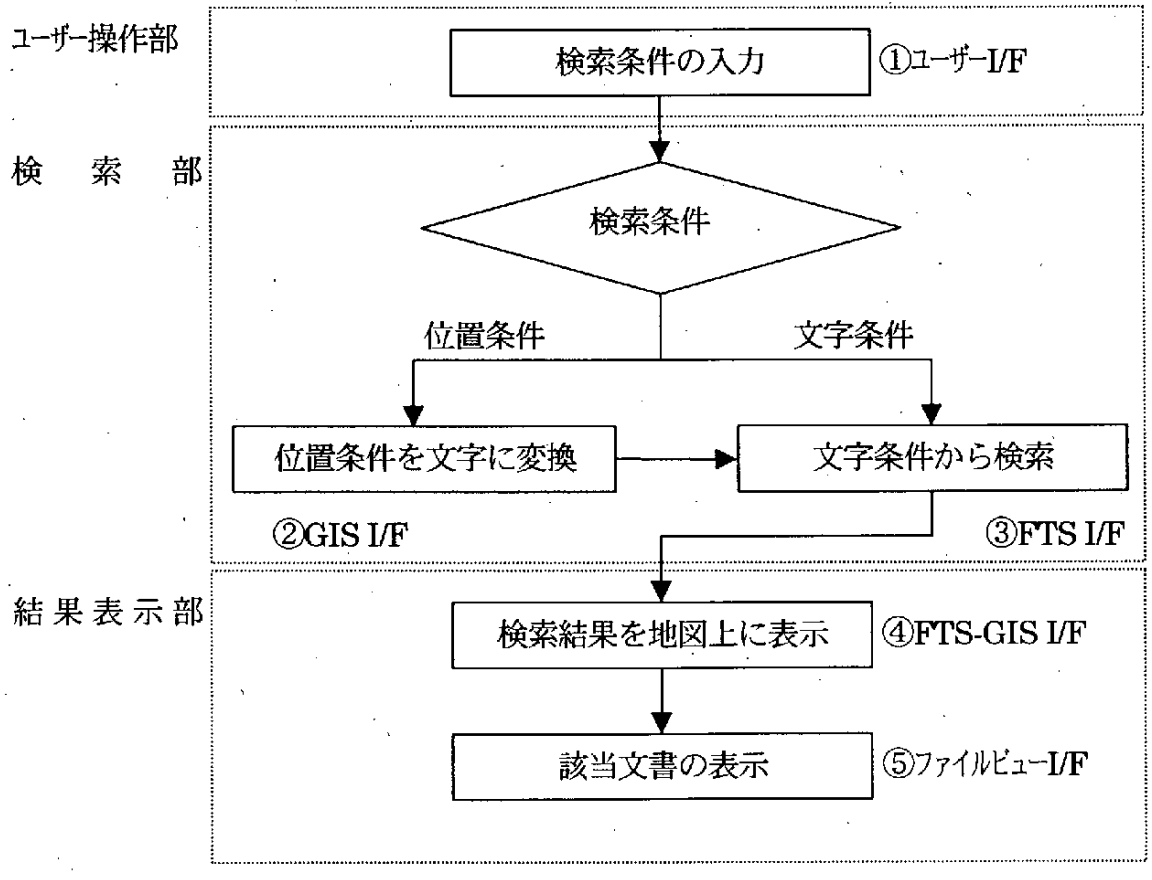

図２－２　フロー図

（2） 各インターフェースの機能

①ユーザー I/F

検索の受付画面となる。検索条件に応じた画面を表示し、検索対象となるDB受付部に処理を渡す。

②GIS I/F

検索条件が地図上の位置に関するものの場合、そのままでは全文検索を行うことが出来ないため、位置条件を文字条件に変換する。

③FTS I/F

ユーザー I/F またはGIS I/F から渡される文字条件をもとに、全文検索サーバーに検索要求を発行する。サーバーからの検索結果を受け取り、FTS-GIS I/F に処理を渡す。

④FTS-GIS I/F

検索結果を地図上に表示する。FTS I/F から受け取った位置情報と文書情報を解析し、該当する文書情報を地図上に展開する。

⑤ファイルビュー I/F

FTS-GIS I/F により文書情報が表示されたのち、ユーザーの要求により詳細を表示する。文書情報をもとに該当する文書のファイル形式を確認し、該当するアプリケーションを起動する。

（3） その他の機能

①地図表示の切り替え

地図の拡大・縮小や、結果の印刷機能等は、ベースとしたアプリケーションが持つ機能を用いる。

②管理ツール

全文検索のメンテナンスや、インデックスの更新を行うツールである。

2．3 開発手法と主要技術の整理

本システムを作成するために必要とされる主要技術の整理を行う。

2．3．1 地理情報技術

GISは、文字や数字、画像などを地図と結びつけて、画面上に地図を表示して、位置や場所に関連するさまざまな情報を統合、分析し、分かりやすく表現することができる仕組みである。行政や市民生活やビジネスの現場で幅広く活用でき、今後利用拡大が期待されている。

現在、多くのGISソフトが市販されている。以下に、主なGISソフトについて簡単にまとめる。

表2－1　GIS製品

製品名	主な特徴
MapInfo	ｶｽﾀﾏｲｽﾞ：専用ｶｽﾀﾏｲｽﾞﾂｰﾙ（MapBasic）、VisualBasic等 その他特徴：利用できる地図は多いが事前に変換が必要
SIS	ｶｽﾀﾏｲｽﾞ：VisualBasic等 その他特徴：変換無しに利用できる地図が豊富
ArcView	ｶｽﾀﾏｲｽﾞ：専用ｶｽﾀﾏｲｽﾞﾂｰﾙ（Avenue） その他特徴：最大シェアを持つESRIの製品
AutoCADMap	ｶｽﾀﾏｲｽﾞ：専用ｶｽﾀﾏｲｽﾞﾂｰﾙ（ObjectARX）、VisualBasic等 その他特徴：CADベースであるため、地図作成は得意

それぞれのGISソフトは単体で利用することができるが、用途により必要とする機能、データ、地図は様々であるため、実際の利用には何らかのカスタマイズが必要となる。このため、GISソフトにはカスタマイズ機能を備えたものが多い。

本システムは、GISと全文検索を連携したシステムであるため、GISソフトをそのまま利用するのではなく、カスタマイズが必要となる。

GISを利用するにはGISソフトと共に地図データも不可欠である。

地図データは作成にコストがかかるため、まだまだ高価であるが、国土地理院の数値地図2500をはじめ安価な地図も整備されつつある。

地図に含まれている図形データ、属性データは様々であり、多くの地図では、基本的な地名、公共の建物等の情報が含まれているが、利用用途に応じた必要な情報が含まれていないければ、別途データを投入する必要が発生する。

以下に、市販されている地図データの一部を簡単にまとめる。

表2－2　地図の種類

地図名称	販売元、特徴
数値地図2500 （空間データ基盤）	販売元：国土地理院 安価に入手できる、ﾃﾞｰﾀﾌｫｰﾏｯﾄが公開されている。
ダイケイマップ	販売元：ダイケイ 都市計画図ベースにしている。
ZMapTownII	販売元：ゼンリン 住宅地図の電子版、属性情報が豊富。
GISMAP2500V	販売元：北海道地図 属性情報も多く、ﾌｫｰﾏｯﾄが公開されている。

GISを利用して、活用したい資料と地図をあわせて活用するには、GIS上の地図と資料を関連付ける作業が必要である。

活用したい資料には、表や台帳のように項目が定まっている定型情報と、書式の定まっていないワープロ文書のような非定型情報がある。活用したい資料が統一されたフォーマットで定型情報になっていれば、比較的容易に地図との関連付けを行って利用することができる。しかし、利用したい資料が非定型情報である場合、GISが利用しているデータベースは一般的にRDB（リレーショナルデータベース）であるため、正規化を行いRDBで扱える形に変換する作業が発生する。

２．３．２ 全文検索技術

全文検索技術とは、文書内のすべての文字列を、直接、検索対象とすることが可能な検索方法である。

（1） 普及の要因

全文検索技術の歴史は比較的長いが、近年複数の要素により一般化しつつある。その要因として以下の項目が挙げられる。

オフィス文書の増加

- 業務におけるコンピュータの利用比率が高くなり、文書をサーバーに大量に作成するようになった。しかし、多くの場合では文書を保存するための管理であり、文書を活用するという視点が欠けている。このような文書は死んだ情報となりやすい。整理・保存された文書より、雑多な整理されていない文書から、そこに埋もれている必要な情報を活用できるかどうかが重要な要素となってきた。
- オフィスに埋もれている情報を積極的に利用しようとする動きからナレッジマネージメントやデータマイニングといった手法が注目されはじめた。

インターネット

- インターネットの普及により全文検索システムの利用者が増加した。
- WWW により電子化された情報発信が容易となり、データが手作業では検索不可能なほどに増加した。

（2） 全文検索の特徴

オフィス文書の増加やインターネットにより、全文検索技術に対する需要やニーズが増加し、製品数の増加、性能の向上が図られた。全文検索技術が従来のデータベースと異なるのは、文書そのものを対象とするところである。全文検索技術の特徴を以下にまとめる。

従来の検索の問題点

- 文書の検索にはキーワード追加が不可欠で、その追加コストが大きい。
- 検索式が複雑になりがちであり、検索結果の質が検索者の力量により異なる場合が多い。
- シソーラスが活用できないため、人力に頼る場合が多い。

全文検索の長所

- 従来の検索システムで必要であったキーワード追加が不要である。
- メンテナンスが原則的に不要となる。
- シソーラスの利用により、類似項目の抽出が行いやすくなる。

全文検索の短所

- 不必要な情報も拾ってしまうため、適合率が低下しやすい。
- データの冗長化が起こりやすく、インデックスが肥大化する。
- 一度登録したものを削除することが困難。

（3） 全文検索技術の手法

全文検索技術にはいくつかの手法がある。以下に例を示す。なお、現在利用されているシステムについては参考文献４を参照して頂きたい。

全文検索技術の例

・N文字インデックス

入力されたテキストをN文字ごとに区切り、各文字列が含まれる文書の情報と文字列の位置情報によるインデックスを作成する。このインデックスを検索し、文字列間の位置関係をチェックして該当する文書を抽出する。

・形態素解析

日本語を解析するための辞書を使用して、入力されたテキストを単語に分解し、この中から名詞などのキーワードを抽出してインデックスを作成する。このインデックスを検索することにより該当する文書を抽出する。

２．３．３ 利用技術の選定

文書に含まれた地理的情報を GIS と連携して利用するためには次の技術が必要となる。

（１） RDB と同様なインターフェースを持つ全文検索システム

全文検索システムは有償無償含めて多くの製品があるが、大部分の製品は検索技術の違いに関わらず Web ブラウザをインターフェースとしたインターネット・イントラネットでの利用を想定している製品が多い。本システムにおいても Web ブラウザでの利用は考慮すべきではあるが、扱うデータが手元にある文書ファイル類であること、GIS 機能を利用することを考慮するとスタンドアローンまたはクライアント／サーバ型アプリケーションで十分である。

また、GIS と全文検索との連携が必要で、プログラムとのインターフェースが重要であり、また住所情報等をデータベースに格納することを考えると一般的な RDB 機能も利用できるほうが望ましい。

これらの条件から、プロトタイプの作成にあたっては、RDB をベースとして、N 文字インデックスによる全文検索機能を持つ Oracle8i interMediaText を利用することとした。Oracle8i interMediaText は全文検索エンジンの機能も Oracle と同様に SQL インターフェースを利用できるためプログラムから利用しやすい。

（２） 検索結果や操作を視覚的に行える GIS ソフト

一般に市販されている GIS ソフトは実装方法が異なるものの、カスタマイズ機能は備わっている。プロトタイプ作成にあたっては、単に GIS ソフトのカスタマイズだけではなく、全文検索システムとの連携が含まれているため、GIS ソフトに付属するカスタマイズ環境よりも汎用的な開発環境の利用が便利である。

また、利用地図についても、市販されている地図を用いることになるが、検索対象となる文書によって利用すべき地図も異なるため、対応する地図のフォーマットも豊富であるほうが良い。

これらの条件から、プロトタイプの作成にあたっては、汎用的な開発環境である VisualBasic で柔軟にカスタマイズでき、各市販地図のフォーマット対応も充実している SIS を用いることとする。

（3） 文書に含まれる地名情報の抽出技術

全文検索システムと GIS ソフトを用いることによって、文書中の情報を検索し、その結果を地図上に反映させることは可能になるが、文書中に含まれる地名情報を抽出して、地図上にプロットすることができない。これを実現するには、文書から地名情報を抽出、地名情報とそれに対応する位置座標とのマッチングが必要となる。地名情報の抽出については、文書の形態素解析を行い、地名と位置座標のマッチングについては、予め作成した地名と座標のデータベースを参照することで実現できる。

これら一連の処理を行うソフトは一般に流通していないため、このプロトタイプにおいては、東京大学空間情報科学研究センターにて研究中である「芭蕉」システム（参考文献１）を利用することとする。

（4） システムの拡張や別システムへの応用を考慮した開発環境

プロトタイプとして作成したシステムの拡張や、内部機能を他システムへ応用することを考えた場合、システムを構成する機能はモジュール毎に分けて作成する。また、開発環境については、各ソフトとの連携や開発の容易さから VisualBasic を用いることとする。

3 プロトタイプの構成と仕様

3．1 プロトタイプ構成

開発するプロトタイプはシステム全体のうち、文書に含まれている地理情報と地図を関連付けて検索、表示を行う部分となる。

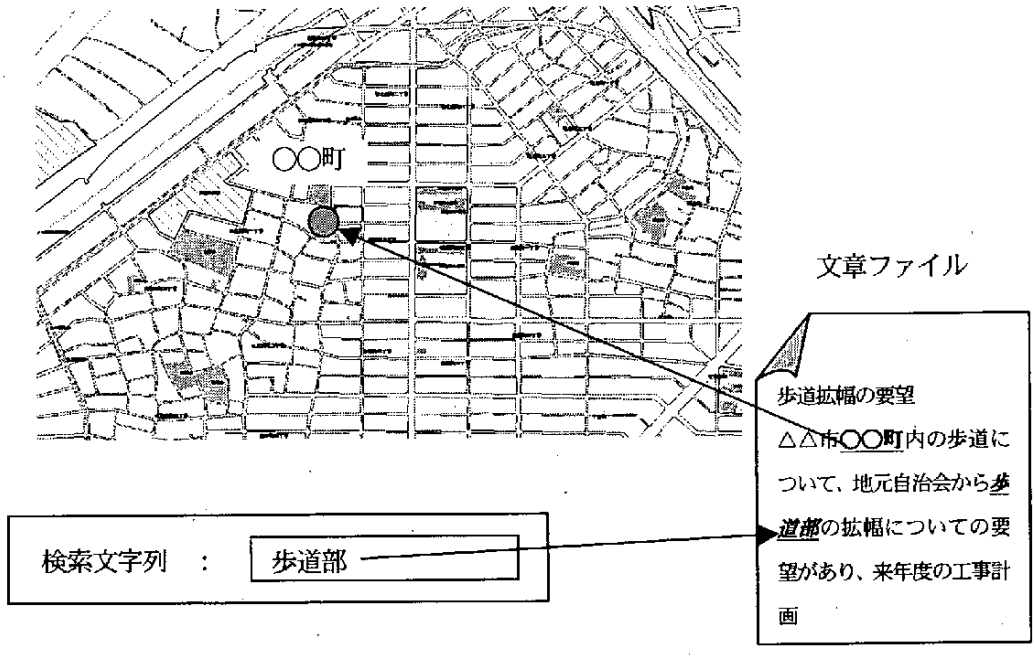

図３－１　プロトタイプの概要

システム全体は図１－２　システム全体構成にあるように文書ファイルだけでなく紙文書の電子化から蓄積までを含むクライアント／サーバシステムで構成されるが、プロトタイプでは電子ファイルのみを対象とし、蓄積文書と地理情報を結びつける部分以外はスタンドアローンで動作するように構築した。(但し、全体としての稼動を見据えてクライアント／サーバシステムで運用することを考慮する)

このプロトタイプは、主に３つの機能に分けることができる。それぞれの機能は、

①地図表示とユーザーインターフェース
②文書のインデックス作成と全文検索
③住所情報抽出と位置座標関連付け

となり、各機能は以下のような関係となる。

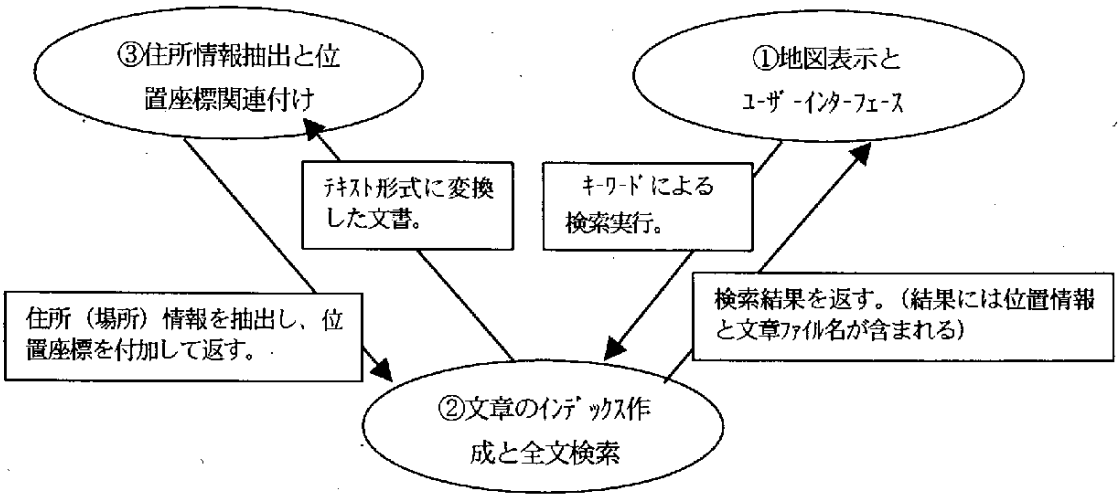

図３－２　各機能の関係

検索対象とする文書については、「文書のインデックス作成と全文検索」機能で管理し、インデックス作成時と同時に位置情報を「住所情報抽出と位置座標関連付け」機能を呼び出して実行する。

利用者は「地図表示とユーザーインターフェース」機能を利用して、文書の検索と結果表示を行う。

3．2 プロトタイプの仕様

3．2．1 「地図表示とユーザーイ ンターフェース」の機能

地図表示とユーザーインターフェースは利用者とシステムとの接点となり、利用者からの要求入力や結果表示を行う。利用者にとってはこの機能の操作のみで文書の検索を行うことができる。

以下に必要とされる機能を挙げる。

・地図が表示され、住所、キーワードによる検索を行う。

　検索を行うための住所選択画面やキーワードの入力画面によりユーザーが作業を行えるようにする。

・任意の住所に関する文書を検索し、地図上の該当地点にプロットする。

　ユーザーから入力されたリクエストに対する結果を画面に表示して、該当する文書を表示できるようにする。

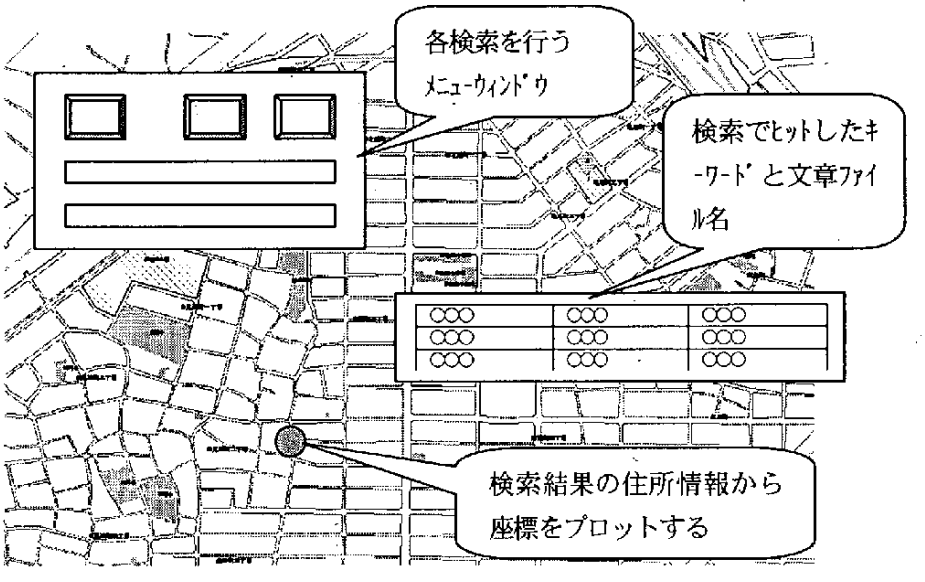

図３－３　地図表示とユーザーインターフェースの概念図

3．2．2 「文書のインデックス作成と全文検索」の機能

検索対象となる文書に対して、全文検索エンジンによる検索と文書のテキスト化を行う。

特定のディレクトリに格納された文書は、全文検索エンジンによりインデックス化されると共に、位置情報抽出のためテキストデータとして保存される。一連の作業は全文検索エンジンによって実行される。

一般的な業務において取り扱う文書は、汎用的なテキスト形式よりもワープロソフト等の特定のアプリケーションにより作成されており、データ形式はそれぞれ異なる。従って、各形式の文書を全文検索エンジンにて扱うためにフィルタプログラムを利用して各形式の文書をテキスト形式に変換する。フィルタプログラムは、全文検索エンジンに含まれており、プロトタイプで取り扱う主な文書は次の通りである。

表３－１　検索対象文書形式

形　　式	バージョン
Text 形式	
HTML 形式	
MicrosoftWord 文書形式	Word97,98
RTF(Microsoft Rich Text Format) 形式	
Excel ブック形式	Excel97
一太郎文書形式	Version.8

３．２．３ 「住所情報抽出と位置座標関連付け」の機能

検索対象となる文書と地図を結びつけるには、それぞれの文書に含まれる住所に関連するキーワードを抽出し、対応する位置座標と関連づける必要がある。これを実現させるためにはフィルタプログラムによってテキスト形式に変換された文書から住所文字列を抽出し、位置座標を付加すれば良い。具体的には以下の機能が必要となる。

・文書を解析して、含まれる住所情報を抽出する。
　検索対象となる文書を形態素解析して、住所・地名と該当する単語を抽出する。
・抽出した住所情報に該当する位置座標を関係付ける。
　抽出した単語を、住所情報と位置座標が格納されているデータベースに対して検索を行う。
・住所情報と位置座標をデータベースに格納する。
　検索の結果を住所情報・座標情報・該当する文書名のフィールドを持ったテーブルに格納する。

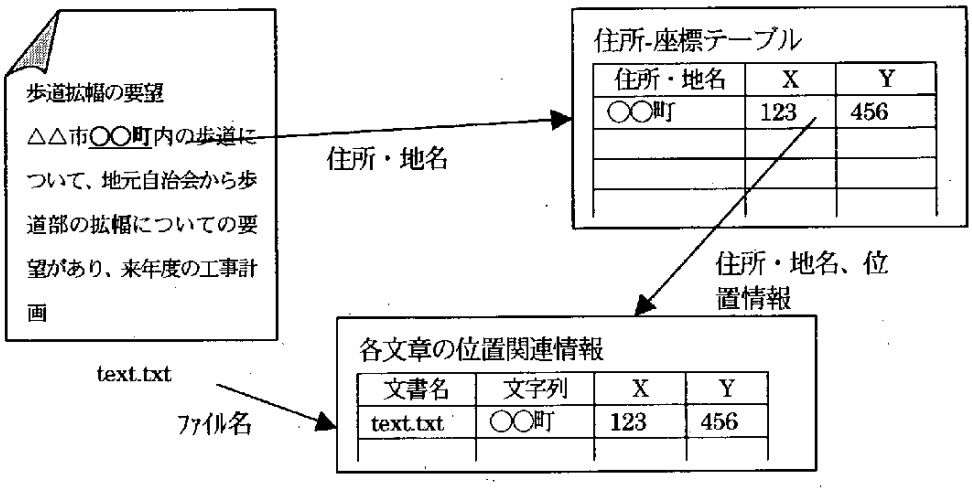

図３－４　抽出された住所情報

3．2．4 利用する地図と住所情報

文書に関連した地図を表示するためには、デジタル地図が必要となる。投入される文書と、利用目的が明確でない場合は、全国地図を準備する必要がある。しかし、自治体や特定の業務、または、限られた範囲で業務を行う事業所での利用を考えた場合、地図はそれぞれの目的に合わせた精度、範囲で整備すれば十分である。

今回はプロトタイプの開発であるため、利用目的による地図の条件は特に制限されないが、画面に表示した時の見やすさと住所データの充実度から Z-MapTownII の愛知県瀬戸市の地図を利用した。

住所データについては、検索対象となる文書中から地名、住所といった空間上の位置特定が可能なキーワードから対応する位置座標を特定するために必要となる。一般的に地名、住所といったキーワードを扱うデータベースは存在するが、これに位置座標が必要になるため、今回はデジタル地図の属性情報を利用した。

キーワードと位置座標の関連付けについては、キーワードとなる住所・地名のデータと位置座標の対応が必要になるため、以下のようなテーブルを持つデータベースを作成した。ここから、住所情報と位置座標の関連付けを行う。

表３－２　住所・位置情報テーブル

項　目	内　容	備　考
拡張市区町村番号	拡張市区町村コード	JIS コード
大字番号	大字コード	
字丁目番号	字丁目コード	町番号
街区番号	街区コード	街区番号
拡張市区町村	市区町村名	
大字	大字名称	
字丁目	字丁目名称	
街区	街区名称	
大字領域	対応する大字の位置座標	19 座標系
字丁目領域	対応する字丁目の位置座標	19 座標系
街区領域	対応する街区の位置座標	19 座標系

3．2．5 処理の流れ

プロトタイプはこれまでに述べた３つの機能をまとめたものとなり、全体の処理の流れは次のようになる。

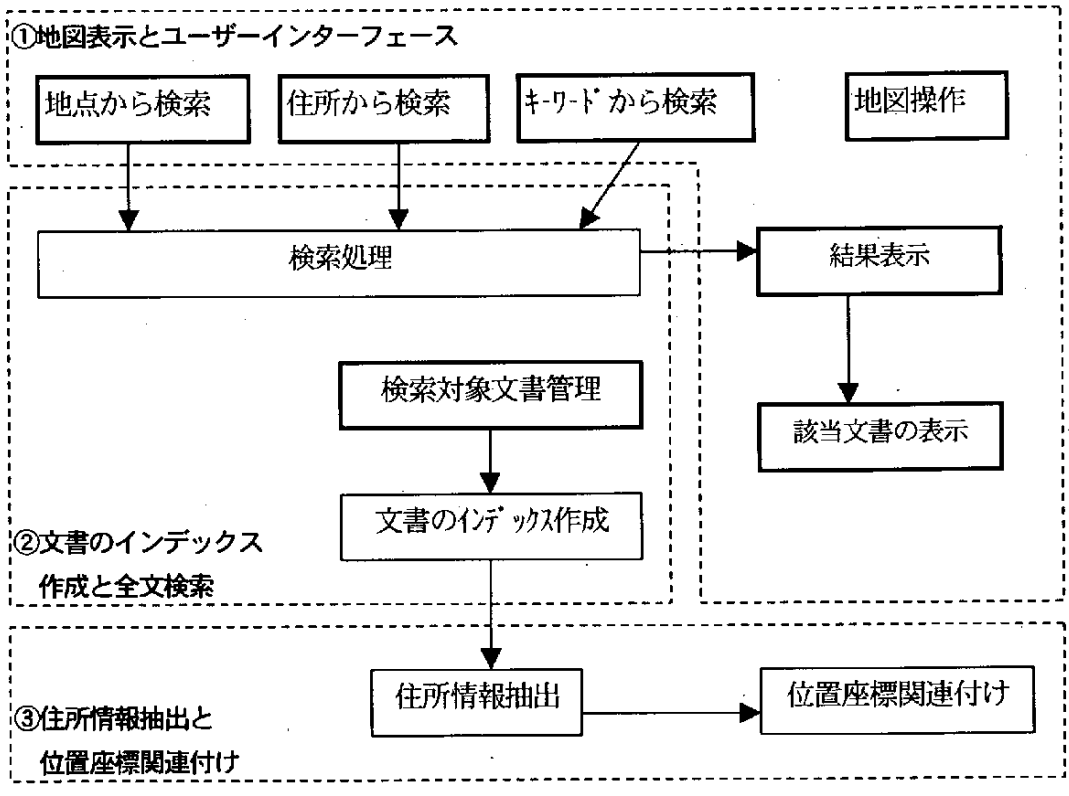

図３－５　処理の流れ

各機能における処理は利用者に対する処理（太線の枠で表示されている処理項目）と、内部で行われる処理に分けられる。

全体的な処理の流れは、利用者による検索と結果の表示、検索対象文書の管理とインデックス作成に分かれる。

3．2．6 データの流れ

プロトタイプで扱う文書データは利用者が検索を行う前の事前作業としてサーバ側で処理される。データの流れは以下の通りである。

①検索対象となるファイルは予め指定されたフォルダ内に格納する。
②全文検索エンジン内のフィルタプログラムにより各形式の文書をテキスト化する。
③テキスト化された文書をインデックス化して全文検索エンジン内に格納する。
④テキスト化された文書から住所、地名に関するキーワードを抽出し、住所データとマッチングを行う住所情報を生成する。
⑤生成された住所情報をデータベースに格納する。

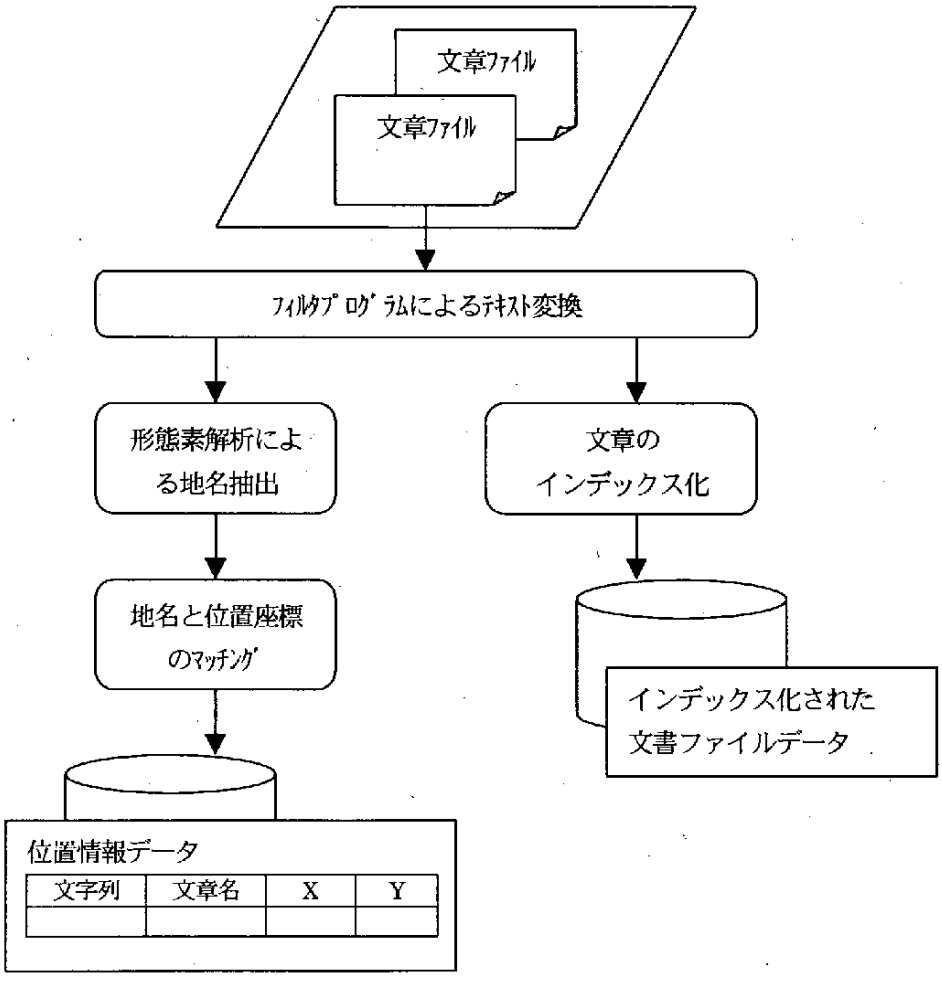

図３－６　文書データの流れ

3．2．7 利用ハードウェア

利用者はWindowsパソコンの使用を想定し、プロトタイプはWindows環境で作成する。

ただし、文書から地名を抽出して位置座標と関連づける機能については利用ソフトの関係で今回はLinuxを用いた。プロトタイプのハードウェア構成を以下に示す。

＜全文検索・ファイルサーバ＞
PC/AT互換機
CPU：PentiumII300MHz
RAM：256MB
OS：WindowsNT4.0Workstation

＜GISユーザクライアント＞
PC/AT互換機
CPU：PentiumII300MHz
RAM：128MB
OS：WindowsNT4.0Workstation

＜住所情報抽出・位置座標設定サーバ＞
PC/AT互換機
CPU：PentiumII300MHz
RAM：128MB
OS：Linux2.2.16

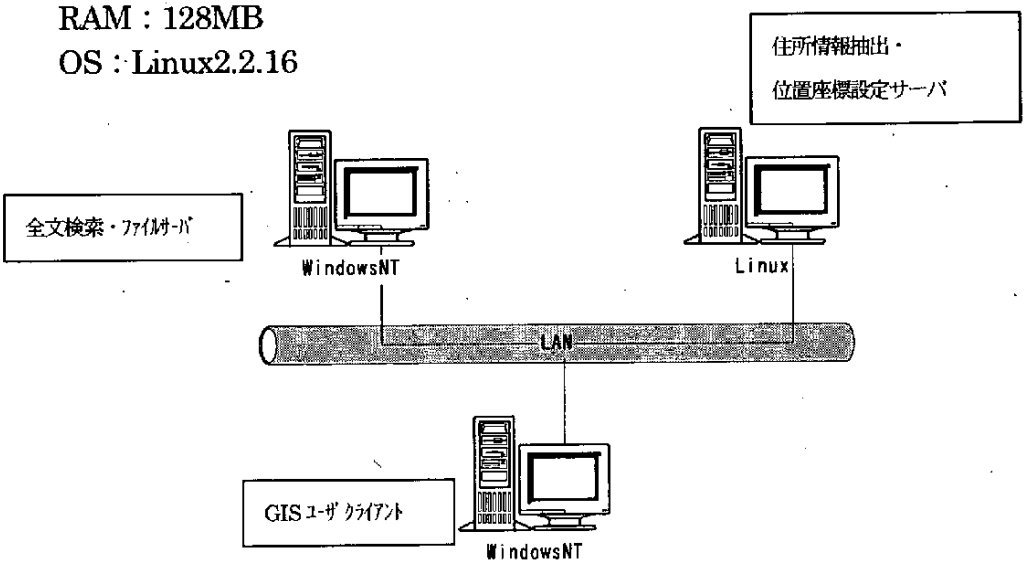

図3－7　機器構成図

3．2．8 利用ソフトウェア

プロトタイプ開発にて利用するソフトと開発言語は、これまでに述べた開発方針に従い以下の構成とした。

地図表示、ユーザーインターフェース機能

・GIS エンジン：SIS5.0

文書のインデックス作成と全文検索

・全文検索エンジン：Oracle8i interMedia

住所情報抽出と位置座標関連付け

・文書解析：茶筅 Ver2.2[*1]
・空間情報抽出：芭蕉[*2]

開発言語

利用者が操作する GIS エンジン、全文検索エンジンとの連携を行うユーザーインターフェースの作成

・VisualBasic6.0SP3

---

*1 茶筅：
奈良先端科学技術大学院大学自然言語処理学講座からリリースされた、フリーの日本語形態素解析器である。
詳細は http://chasen.aist-nara.ac.jp/を参照。

*2 芭蕉：
東京大学空間情報科学研究センターにて開発された空間情報抽出システムである。（参考文献１）

4 プロトタイプの利用イメージ

作成したプロトタイプの機能、操作について、実際の画面を用いて説明する。

4．1 プロトタイプの起動

プロトタイプの起動は実行プログラムのアイコンをダブルクリックにて行う。プログラムは検索作業を行う検索ソフトと、文書データのメンテナンスを行う管理ツールに分かれる。

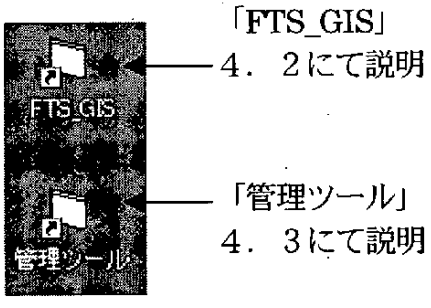

図4－1　プログラムアイコン

４．２ 検索ソフトの機能

４．２．１ 基本画面

検索ソフトは図４－１　プログラムアイコンの「FTS-GIS」を起動することにより基本画面が表示される。基本画面はGISソフトによる地図が表示されたウインドウと各検索機能を利用するためのメニューウインドウ（画面左上）によって構成される。

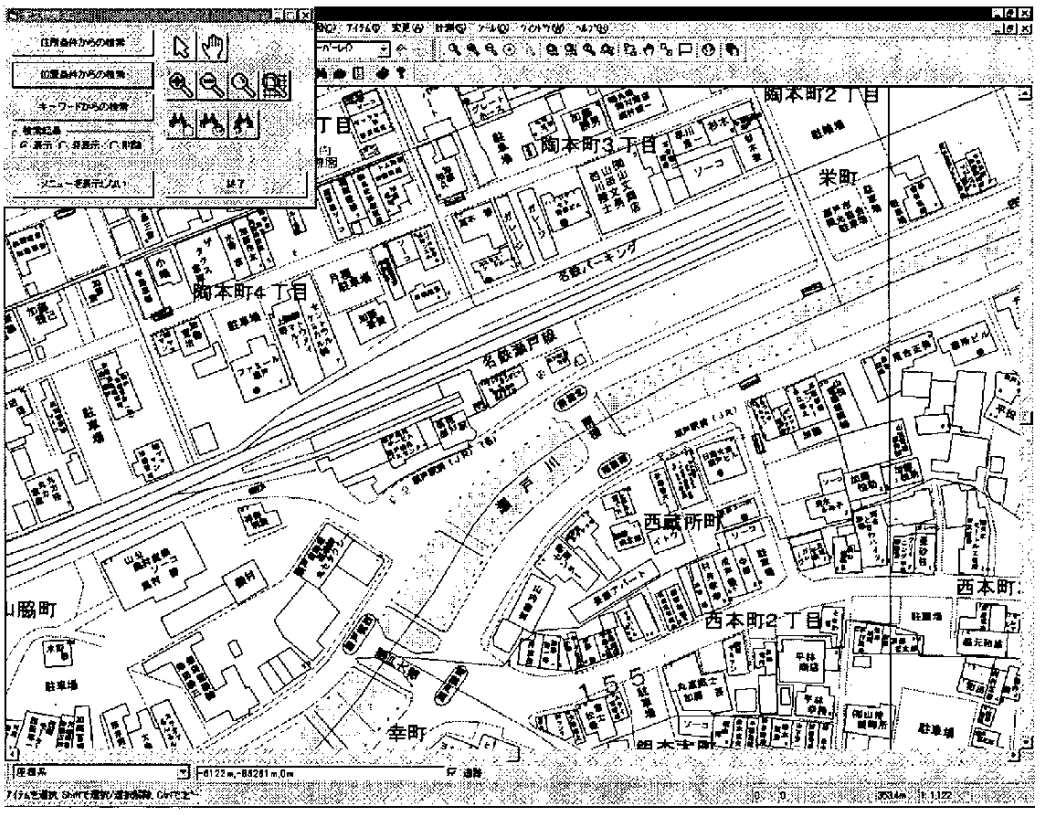

図４－２　基本画面

4．2．2 メニュー画面

検索等の操作は基本的に全てメニューウィンドウに表示されているボタンによって行う。

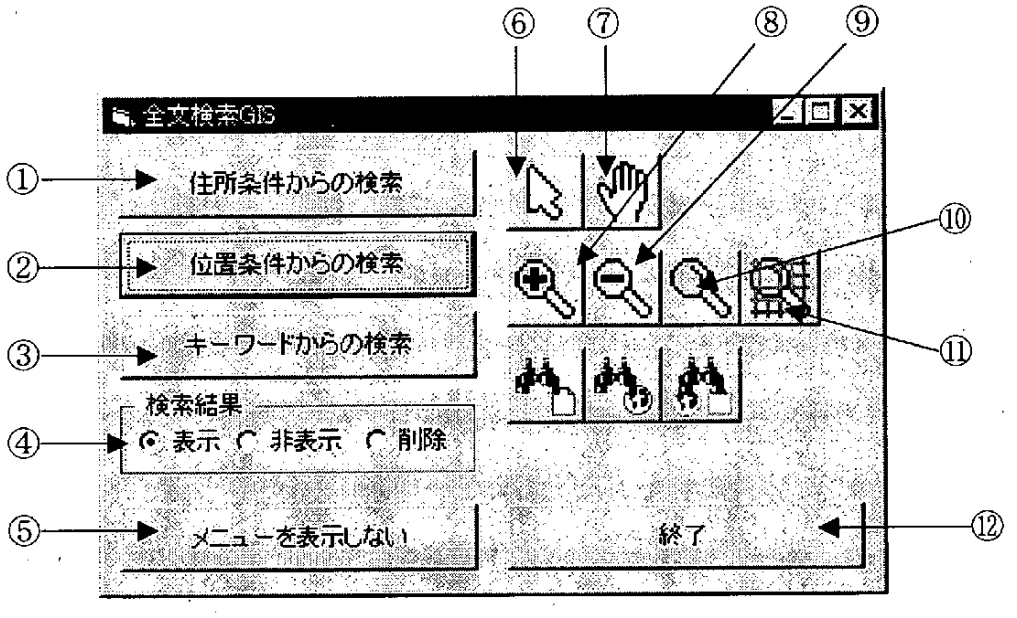

図4－3　メニューウィンドウ

それぞれのボタンの機能は以下の通り。

① 住所の一覧を表示して、選択した住所に関連する文書を検索し、該当する住所の地図を表示します。

② 地図上の任意の点をクリックした場所（住所）に関連する文書を検索します。

③ キーワードに関連する文書を検索します。

④ 検索結果のウィンドウの表示・非表示・削除の変更をします。

⑤ このメニューウィンドウを非表示にします。

⑥ 地図上のオブジェクトを選択できる選択モードに設定します。

⑦　地図の表示位置を移動します（パン機能）

⑧　地図を画面中心に拡大します。

⑨　地図を画面中心に縮小します。

⑩　地図を任意の範囲で拡大します。

⑪　登録されている地図の全範囲を表示します。

⑫　プログラムを終了します。

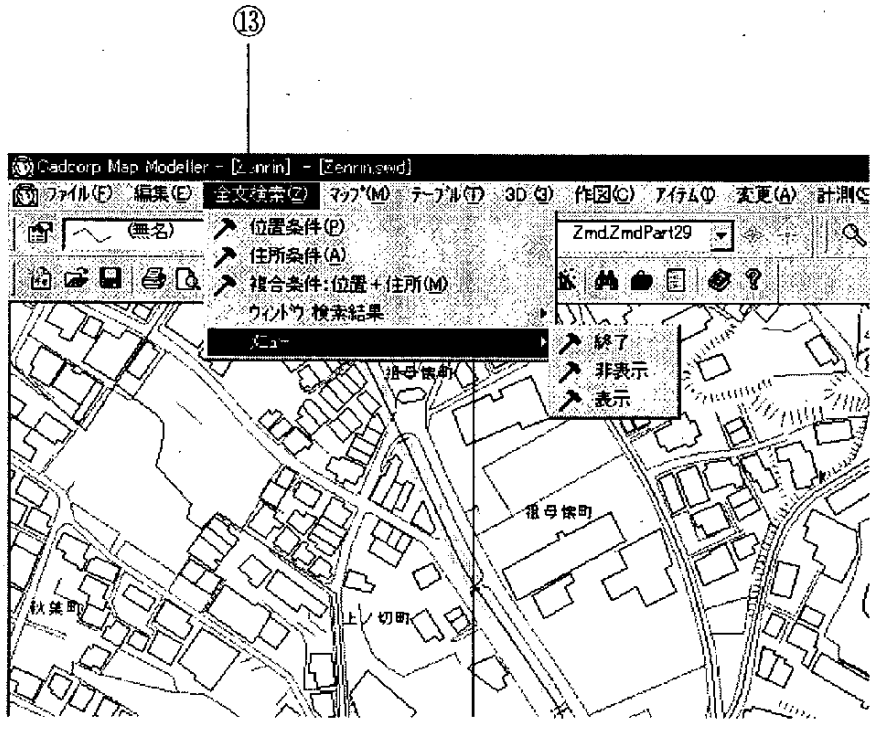

図４－４　メニューバー

⑬　メニューウィンドウの機能と同じものが、GIS ソフトのメニューバー内にも用意されている。

4．2．3 住所指定から検索

一覧表に表示された住所を選択し、その住所に関連する文書の検索を実行する。
利用はメニューウィンドウにある、①の矢印で示されたボタンで開始される。
ボタンが押されると、②の矢印で示された住所選択のダイアログが表示される。

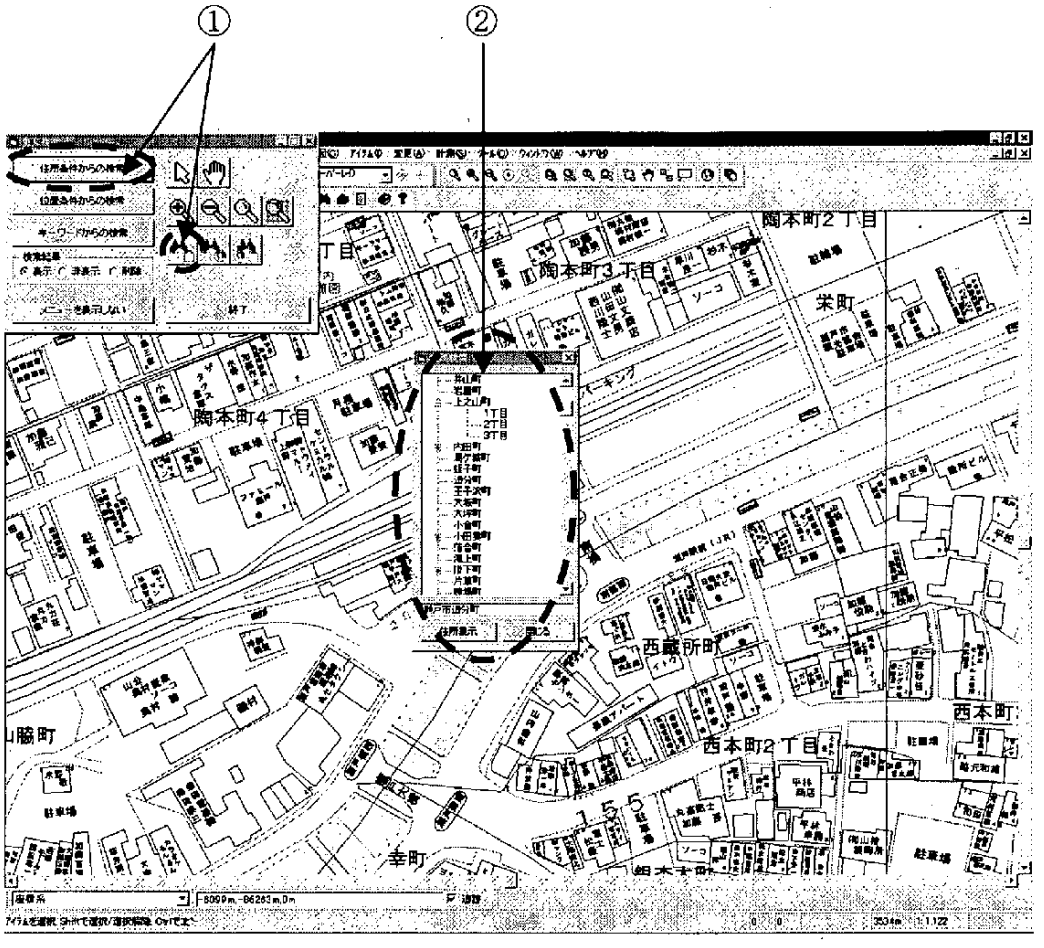

図４－５　住所からの検索

住所選択はツリー構造になっており、市町村名－町名－丁名の順で階層化されている。

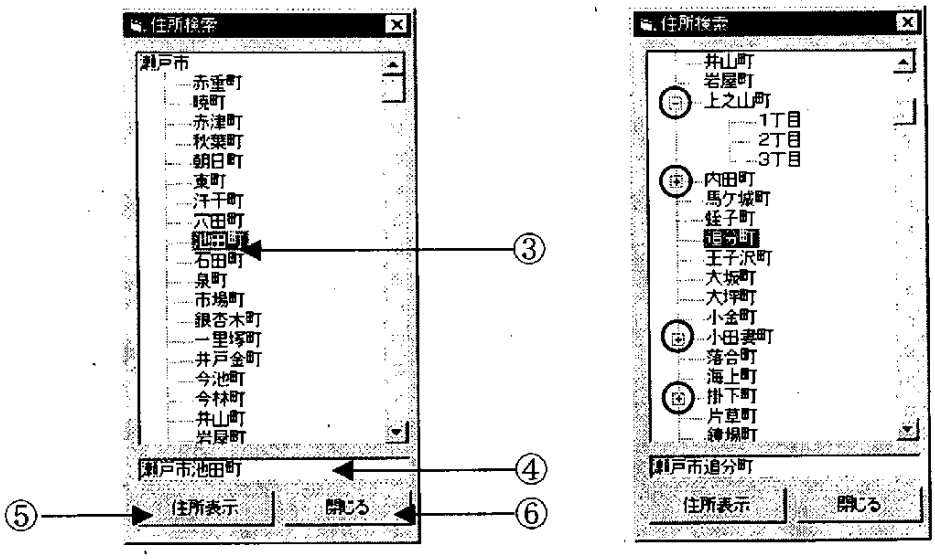

図４－６　住所選択ウィンドウ

③　ツリーリストになっており、下の階層を表示するには「+」の部分をクリックする。下の階層が表示されると、「－」に変化する。
住所を選択すると色が反転し、ダブルクリックで⑤の住所表示ボタンのクリックと同じ動作をする。

④　選択中の住所が表示される。

⑤　住所を選択した状態で押すと、その住所をもとに検索を開始する。

⑥　このウィンドウを閉じる。

４．２．４ 地図上から検索

地図上でクリックした地点での住所の町丁名をキーワードにして関連する文書を検索する。

メニューウィンドウ内の①の矢印で示したボタンをクリックすることにより、地点指定モードになる。この後、地図画面上の任意の地点をクリックすることにより、クリックした地点の住所に関連する文書の検索を開始する。

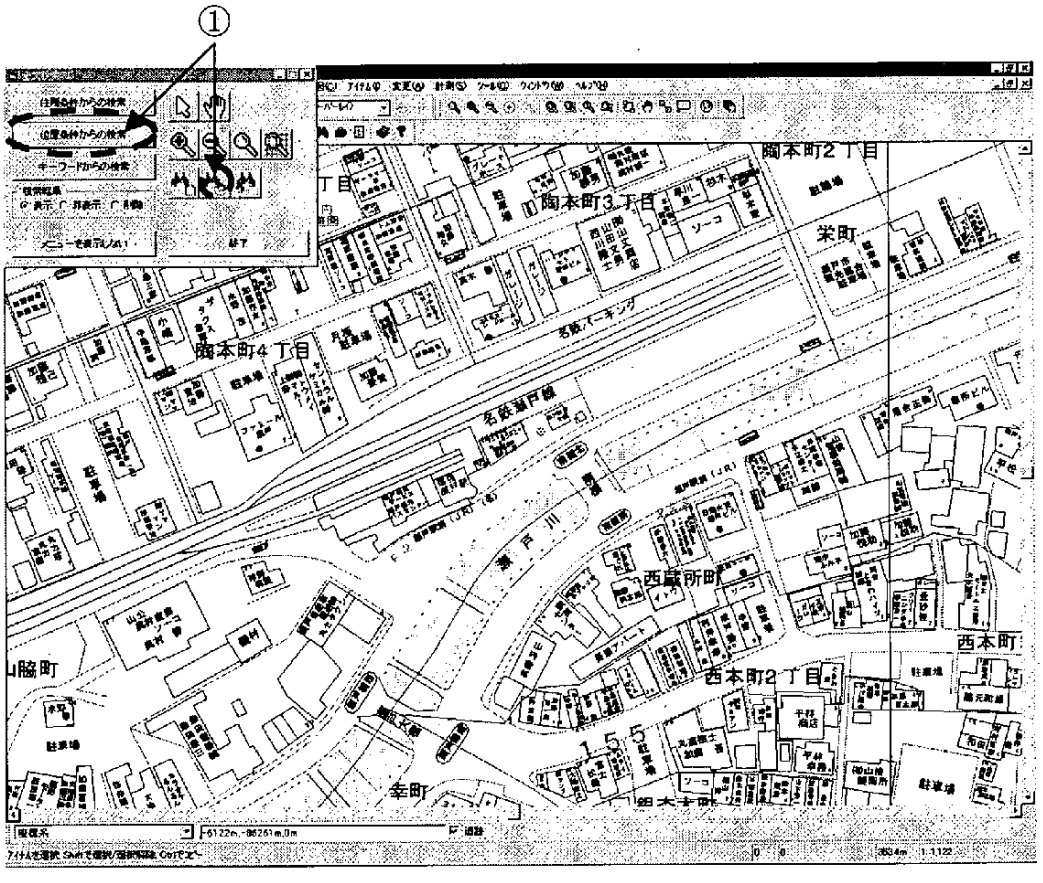

図４－７　地図上から検索

4．2．5 キーワードによる検索

任意のキーワードを入力して関連する文書を検索する。

メニューウィンドウ内の①の矢印で示したボタンをクリックすることにより、「キーワードによる検索」ダイアログが表示される。

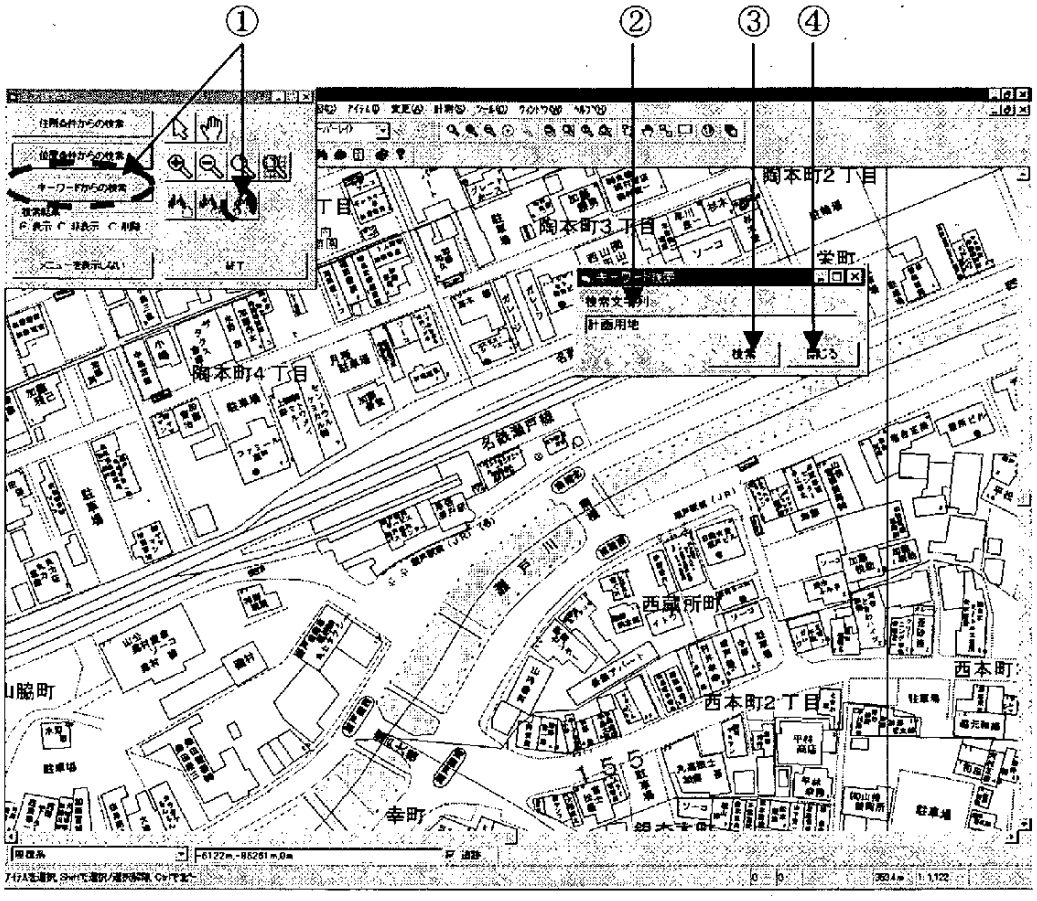

図4－8　キーワード検索

② 検索したいキーワードを入力する。

③ 検索を実行する。

④ ウィンドウを閉じる。

４．２．６ 検索結果の表示

住所指定による検索、地点指定による検索、キーワードによる検索において検索結果の表示は全て同じである。

検索結果については、ヒットした適合率と文書のファイル名がウィンドウに表示される。

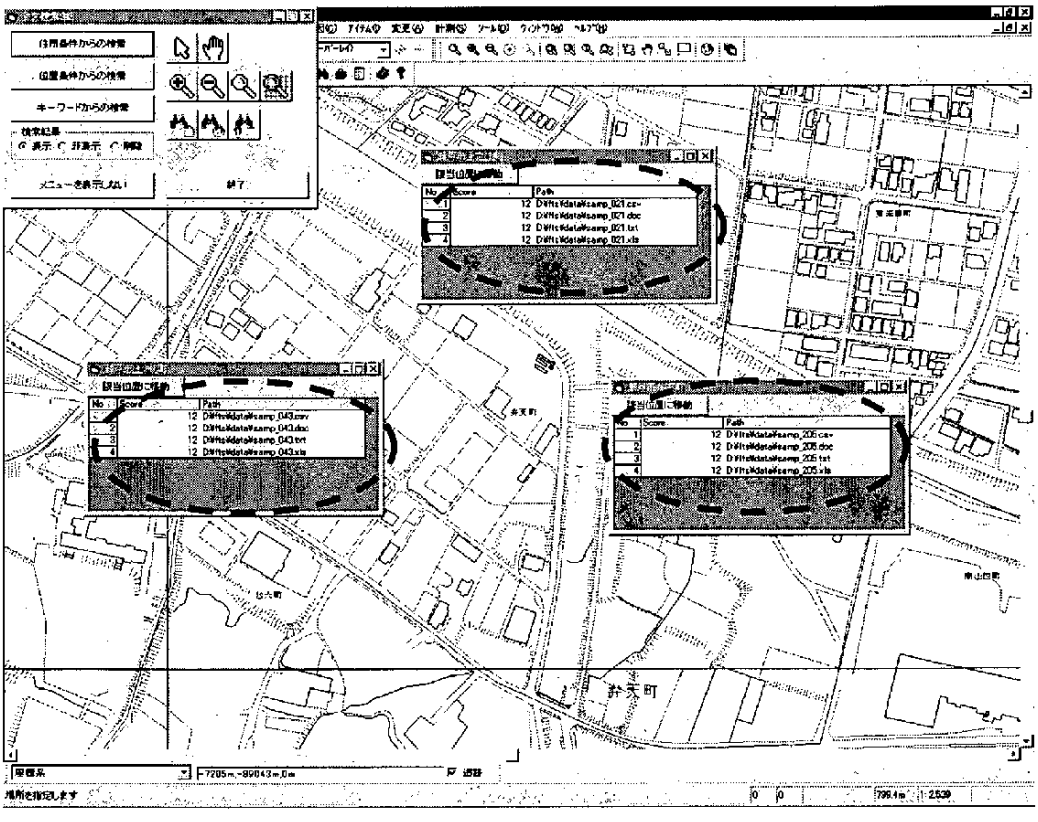

図４－９　検索結果の表示

検索結果は結果表示ウィンドウに一覧表形式で表示される。表示カラムは No（通し番号）、Score（文書の適合率）、Path（文書のファイルがある場所）をそれぞれ表示する。

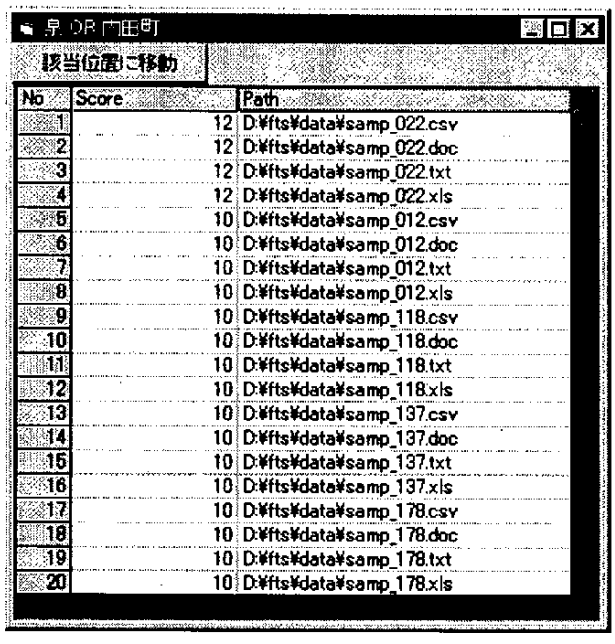

図４－１０　結果表示ウィンドウ

結果表示ウィンドウの各結果項目をダブルクリックすると、該当文書が表示される。

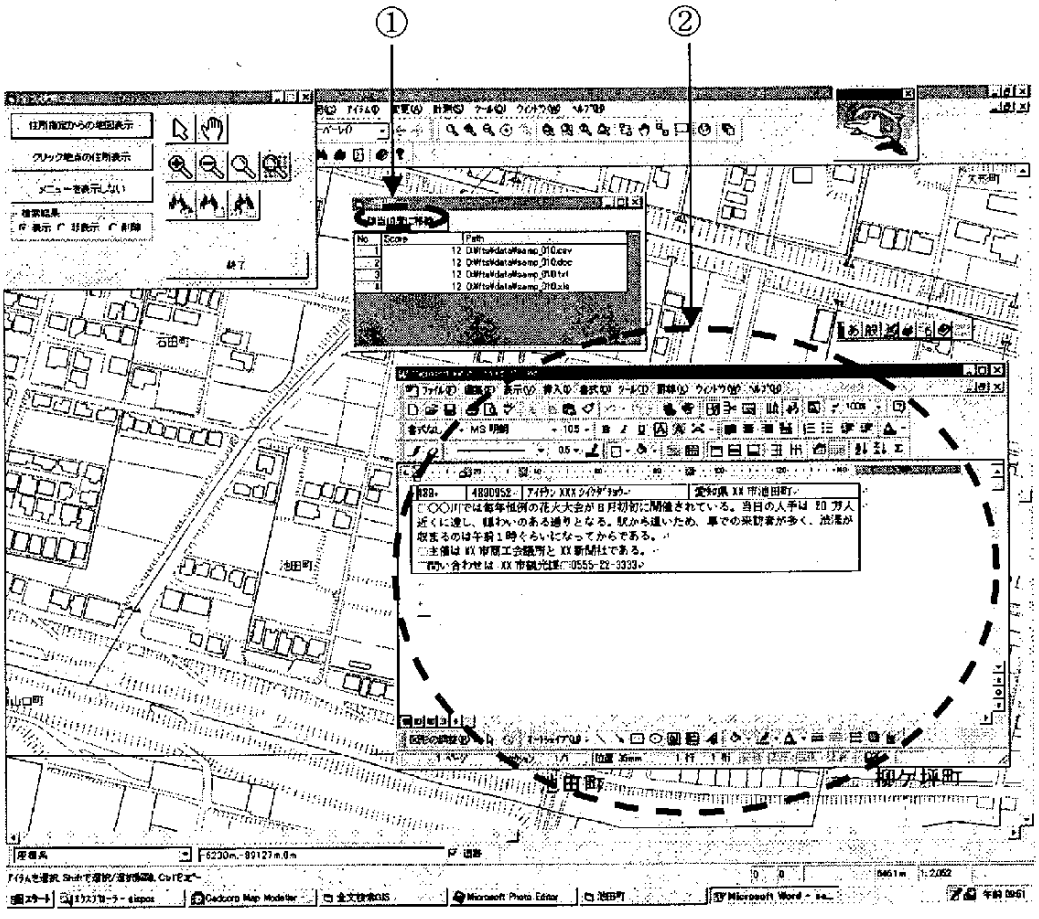

図４－１１　該当文書の表示

① 「該当位置に移動」ボタンを押すと文書に関連のある地点の地図を表示する。

② 関連した文書を表示した様子（Word 文書等）。

４．３ 管理ツールの機能

文書メンテナンスは、検索機能とは別のプログラムで実行される。

図４－１　プログラムアイコンにある、「管理ツール」を起動することにより、「管理者ツール」というタイトルのウィンドウが表示される。

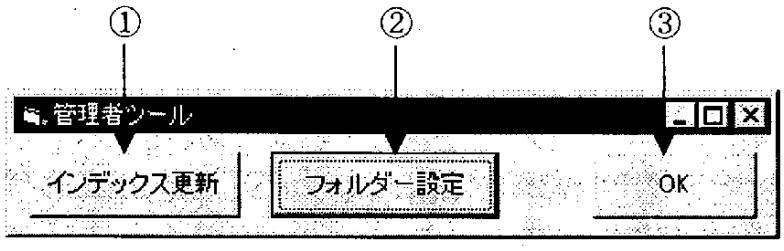

図４－１２　管理者ツールメニュー

① 蓄積された検索対象の文書を検索できるようにインデックスの更新を行う。
インデックス作成は、対象となる文書の数、大きさによって要する時間は異なる。終了時には以下のメッセージボックスを表示して、処理が終了したことを通知する。

図４－１３　インデックス更新終了

② 検索対象とする文書が配置されたフォルダの設定を行う。
(次頁参照)

③ 終了してこのウィンドウを閉じる。

「フォルダ選択」ボタンを押して、検索対象としたい文書を設定する。

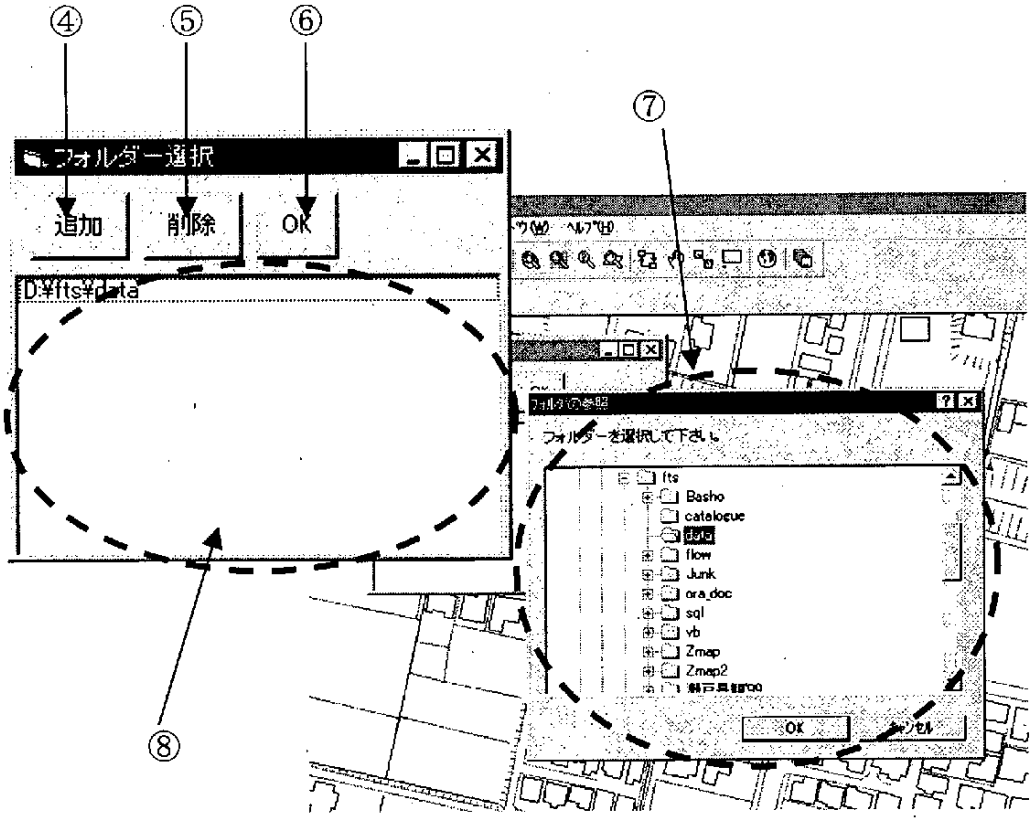

図４－１４　検索対象フォルダの設定

④　「追加」ボタンを押すと④のフォルダ指定のダイアログが開き、検索対象のフォルダを指定する。

⑤　⑧に表示されているパスを選択して、「削除」ボタンを押すと検索対象から外される。

⑥　フォルダー選択ウィンドウを閉じる。

⑦　フォルダ参照ダイアログ。

⑧　検索対象になっているフォルダのリスト。

5 まとめ

5．1 プロトタイプ作成によって得られた効果

プロトタイプ作成により得られた効果は次のようなものである。

①文書の地図上への配置

文書を地図上に配置できるため、対象となる地域と文書の関連が認識しやすくなる。

②２つのデータベースを意識せずに利用できる。

内部的にGISのデータベースと全文検索エンジンのデータベースを用いているが、利用中にユーザーが意識することはなく、GIS が全文検索機能を持ったかのように動作することができる。

③文書と地図の関連付け作業が不要。

従来のGISでは、文書の内容から位置情報を確認し、地図上の位置を登録する必要があったが、今回のシステムでは自動的に登録されるため、関連付けの作業が不要である。

④文書の登録を簡単に行うことができる。

文書の登録は、指定したフォルダにデータを保存することで検索対象となる。特別に内容を登録する必要はなく使用できる。

⑤インデックス作成は簡易な操作で行える。

全文検索のインデックス作成は管理ツールのボタンを押すだけで、作成が可能である。文書の変更・追加・削除はプログラム内部で判断を行い、インデックス内容の変更が行われる。

以上の成果により、当初の目的としていた機能を達成することができ、これにより、GISと全文検索の連携のプロトタイプとして一定の成果を得たと考える。

5．2 今後の展開に向けて

今回のプロトタイプ作成においては、一つの市域内の限られた区域を対象としたデータで、町丁目レベルの住所情報が含まれているものを用いた。今後は、番地レベルの内容を含んでいるものを対象とすることや、より対象範囲を広げることも考えられる。

将来的な展開としては、紙文書をスキャナーで読んだものも対象とし、より包括的な文書管理システムと展開していきたい。さらに、今回のシステム構成ではクライアント／サーバー形式としたが、文書の共有化を進めるためにもウェブ型のシステム構成として展開を図りたい。

今回のシステムで重要なデータとなったのが、地名と位置座標に関するデータである。今回は住居表示まで得られる住宅地図を用いた。今後は、駅名や通称名までを含んだ「地名－位置座標辞書」の整備が望まれる。

今後、全文検索技術のさらなる進歩が見込まれる。新しい技術に対応した高速で正確な連携システムの構築を進め、入力の負担が少ないシステムの構築を進めていきたい。

なお、本システムの開発にあたり東京大学空間情報科学研究センター 助手　相良毅氏より貴重なご示唆を頂いた。ここに謝辞を表わします。

参考文献

1） 相良 毅・有川 正俊・坂内 正夫，「ジオリファレンス情報を用いた空間情報媒介システム」，情処処理学会論文集：データベース，2000

2） 松本 裕治，北内 啓，山下 達雄，平野 善隆，松田 寛，高岡 一馬，浅原 正幸，「日本語形態素解析システム『茶筅』Version2.2.1 使用説明書」，2000，http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

3） 全文検索システム協議会，「全文検索システムとは何か？ 2000 年版」，全文検索システム協議会，2000，http://www.ftsanet.com/ dbtokyo00/Db00.htm

4） 馬場 肇，「日本語全文検索エンジンソフトウェアのリスト」，2000，http://www.kusastro.kyoto-u.ac.jp/~baba/wais/other-system.html

5） 相良 毅・有川 正俊・高橋 昭子，「XML を基本としたテキスト空間情報ベース」，情処処理学会論文集：データベース，1999

6） 相良 毅・有川 正俊・坂内 正夫，「ネットワーク上各種情報源からの地理情報抽出収集手法」，地理情報システム学会講演論文集，1999

平成１３年３月発行

発　行　　財団法人　データベース振興センター
東京都港区新橋２丁目１３番８号
新橋東和ビル５階
TEL　０３－３５８０－２４３０

委託先　　株式会社　創建
愛知県名古屋市熱田区新尾頭１丁目１０番１号
TEL　０５２－６８２－３８４８

印刷所　　中日青写真株式会社
愛知県名古屋市中区新栄３丁目２１番３１号