情　報　学　基　礎　40－7
（1995．11．9）

Information Outlining – 検索情報の可視化 – 行政情報の活用のために

浦本 直彦　諸橋 正幸

日本アイ・ビー・エム株式会社 東京基礎研究所

大量の情報の中から，利用者にとって役立つもののみを簡単に，かつ，的確に取り出す情報検索手法として，我々は，Information Outlining という概念を提案し，それに基づくプロトタイプシステムを作成した．Information Outlining は，ある条件で見つかったデータの集合に対して，個々のデータが持つ属性値（キーワード，書誌情報，など）を複数のビューでクラスタリングし，クラスタを代表する属性値で，データの内容を推測させる手段である．あわせて，このプロトタイプシステムを行政情報の検索・活用のために利用する実験例を紹介する．

Imformation Outlining – Visualization of Extracted Knowledge to Utilize Government Issued Information

Naohiko Uramoto and Masayuki Morohashi

IBM Research, Tokyo Research Laboratory
1623-14, Shimo-turuma, Yamato-shi, Kanagawa 242 Japan
{uramoto,moro}@trl.vnet.ibm.com

We propose an integration of visualization and navigation techniques, called "Information Outlining," for digital libraries. It enables users to catch all the characteristics of the underlying data in different views during their search processes. A prototype system for retrieving government issued information was developed by the "database for new industries/business" project initiated by IPA (Information-technology Promotion Agency).

1 はじめに

大量の情報の中から，利用者にとって役立つものを取り出す情報検索の手法は，キーワードを中心とした，検索キーを与え，該当するデータの件数と各データのタイトルなどを頼りに必要情報を手に入れる，というのが典型的であった．この場合，ほんとうにそれが利用者にとって役立つ情報かどうかの判断は，タイトルなどの簡単な記述にたよるか，実際に取り出したデータを1件ずつ確認するまではわからないことになる．

この問題に対する解決策として，我々は，Information Outlining という概念を提案し，それに基づくプロトタイプシステムを作成した [1] [2]．Information Outlining は，ある条件で見つかったデータの集合に対して，個々のデータが持つ属性値（キーワード，書誌情報，など）を複数のビューでクラスタリングし，クラスタを代表する属性値で，データの内容を推測させる手段である．さらに，各ビューで示されるクラスタは，すべて選択可能（選択することで，そのクラスタに属するデータを検索することになる）として，検索のプロセス（とくに絞り込みプロセス）のどの時点でも，捕まえているデータ集合に対する輪郭が見えているシステムを実現して，これを検索のためのナビゲーション機能と位置づけた．

本年，情報処理振興事業協会 (IPA - Inforamtion-technology Promotion Agency) は，新産業創造データベースセンターを開設し，行政情報に関する文書および統計情報の利用実験を行うこととなった．この行政情報データベースの検索用実験システムの開発に際して，我々の提案していた Information Outlining が採用されたので，ここに IPA の許可を得て，実験システムの紹介をするとともに，その中で活かされている Information Outlining の具体的な応用について述べる．

2 システム構成

システムの概観を図1に示す．

システムにおける主要コンポーネントは，以下の通りである．

- サーバ
  - 検索エンジン
    メニュー検索，全文検索エンジン
  - information outliner

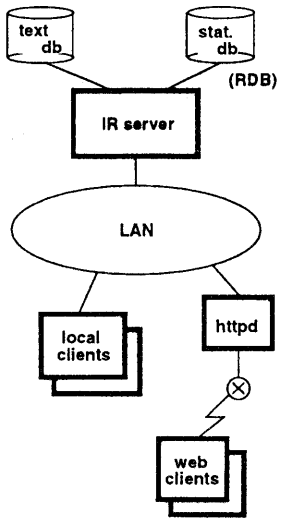

図 1: システム概観図

- 構内あるいは web クライアント
  - 検索 GUI
  - 検索された情報内容のビューワ

2.1 サーバの機能

サーバは検索エンジンと information outliner からなる．検索対象となるデータは，通常の情報検索システムで扱うテキスト中心のデータ（以後，文書データと呼ぶ）と，ＲＤＢに格納される数値データ（統計情報を中心とするデータ）があるが，検索エンジンが直接扱うのは文書データのみである．数値データに関しては，格納された表ごとに，それらのメタ情報を文書データとして用意することで，ユーザは，間接的にこれらの表の存在を知り，ＲＤＢ内の表へのアクセスが可能になる．

文書データにもメタ情報を用意することは可能である．文書データに対するメタ情報は，図書館システムなどで使われている書誌データとして位置づけることができる．本システムでは，文書データに関しては，文書そのものとその書誌データに区別を設けないので，両者（データベース中に書誌データとその文書の全文データの両方が存在していれば）が同時に検索可能となる．このことは，見かけ上，検索件数を最大2倍に増やすことにな

るが，検索結果に無駄なデータを含まないところまで絞り込む検索機能が備えられたシステムならば，邪魔にならないどころか，対象データに対する熟知度によって，書誌データを見たあとで実際の文書を読んだり，いきなり，実際の文書を見に行ったりという選択が可能になるので，かえって都合がよい．

従来，全文データ（または抄録データ）を対象とする情報検索システムと，書誌データを対象とする図書館システムとは，別個のシステムと考えられてきた．本システムでは，それらを1つのデータ（全文データとみなして）として扱う．その場合に問題となるのは，書誌データの持つ，タイトル，著者，出版社，等々といった構造をどう保存するかという点である．そこで，我々は，書誌データをSGMLフォーマットで表現する方式を導入した．これにより，検索において，全文データ向きの検索方法と，書誌データ向き（通常のDBシステムに近い）検索方法の両方が可能なシステムを実現した．また，SGMLの導入により，SGML文書が持つ本来の利点であるマルチメディア文書の保管が可能になった．なお，これらの利点については，後にもう一度論ずることにする．

システムにおける検索エンジンは，必ずしも1つに限定する必要はない．それぞれの検索エンジンがどのような能力を持つかが，はっきりとユーザに分かるならば，複数エンジンを持つ検索システムは，ユーザにとり，能力の高いシステムとしての評価を受けるであろう．現時点で，本システムが持つ検索エンジンは，メニュー検索と全文検索の2つである．前者は書誌データのような整理された属性検索に適した検索エンジン（「3.4 マッパ」参照）であり，後者はいわゆる「情報検索システム」に多用される検索エンジンである [3]．

数値，文書にかかわらず，データベースから必要な情報を取り出すときに，ユーザが意図するデータがきちんと取れることが必須条件であることはいうまでもないが，実際に検索するにあたっては，取り出したデータの量が一定時間内で理解可能な程度に少ないこともまた，重要な要素である．したがって，検索のプロセスには，該当データの量が，十分少なくなるまで絞り込むという作業が必須となる．本論の冒頭でも述べたように，この絞り込みの過程で，重要なデータを落としていないか，あるいは，絞り込みに用いる追加の検索条件が適切であるかどうか，といった確認がとれるような仕組みを実現するのがinformation outlinerである．具体的な機能とその実現法は後述する．

2.2 クライアントの機能

クライアントにおける機能設計の主要な考慮点は，以下の2つである．

- Information outliningによる検索の途中経過や最終結果をなるべく図示化すること．図示化されたオブジェクトは，選択可能 (clickable) にすること．
- 構内ユーザとwebユーザがなるべく同じGUIで同じ機能を享受できるようにすること．

この方針に則って，システムがどう実現されたかについては，後述する．

通常の検索システムと異なり，本システムでは検索されたデータ（含メタデータ）の内容はサーバによって供給されず，検索クライアントシステムが自分でアクセスしに行く．ただし，webクライアントに対しては，httpデーモン内の検索クライアントプログラムが，webクライアントに代わってデータのアクセスに行く．

3 Information Outlining Paradigm

information outliningのパラダイムは，図2の形で表現される．

3.1 data-model-viewerの動的関連

何らかの形で選択されたデータ集合に対し，その中のある属性（または属性の組合せ）に従って定義されるモデルを設定する．このモデルをユーザに分かりやすいGUIを通して見せるのがviewerである．information outliningにおける特徴は，data-model-viewer間の対応が両方向であることである．

すなわち，システムは，データ集合からモデルを表現するのに十分な情報（属性とその値）を抽出し，そのモデルにしたがってGUIによる図示を行うという上から下への処理の流れをすべてのモデルやビューに対して実行すると同時に，ビューワから属性値（複数も可）を選択することで，モデルを介してその制約に合致するデータ集合のみ

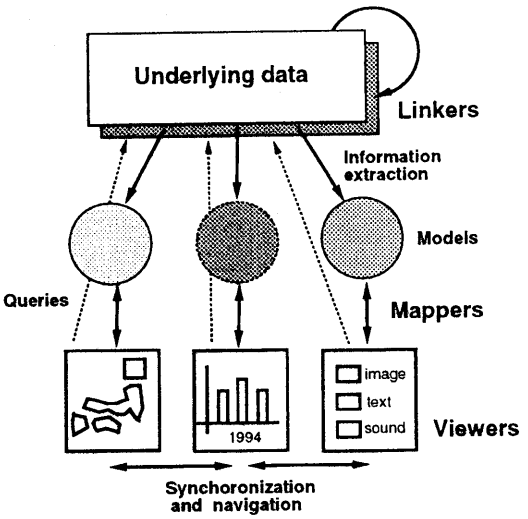

図 2: Information Outlining Paradigm

を残すという下から上への処理も行う．検索における絞り込みの過程で，常に，この上下の処理を両方向にわたって行うことで，ユーザが注目しているデータ集合の性質が，ビューワを通し，様々な観点から，常時，把握できるようになる．

3.2 リンカ

リンカは，個々のデータ間の関連をつけるための機能で，HTMLやSGMLに見られるマルチメディアデータのハイパーリンクなどの静的なリンク形式がその代表であり，本システムでもこの機能を利用している．加えて，本システムでは，メタデータと原データとの関係（書誌と全文，RDBのメタと表の関係）も把握しており，必要に応じてリンカがこの対応を調べ，必要な方をビューワに表示することができる．

3.3 モデルの定義と情報の自動抽出

現在，implementされているモデルは，単一属性からなる単純な形のモデルなので，図3のような形式で属性値を並べ上げて定義することでカスタマイズ可能にしてある．なお，属性値を階層化（図中のピリオドは階層レベルの深さを示している）することで，該当件数の数え上げは，当該属性値に関するものだけでなく，その下位に位置する値のものを含めた形でも数えられるようになっている．

```
日本
. 北海道
. 東北
. . 青森
. . 岩手
. . 宮城
. . 秋田
. . 山形
. . 福島
. 関東
. . 茨城
      :
      :
```

図 3: 属性「日本の地名」によるモデルの定義例（一部省略）

データとモデルの間の対応は，属性値がデータあるいはメタデータ中に含まれるかどうかで判定する．発行年などのように，書誌データ中の特定のフィールドに置かれていることが保証されている場合は，そのフィールド（SGMLのタグにより分かる）に書かれている値で対応を判断するが，それ以外の場合は，テキスト全体の中から属性値に対応するキーワードが使用されているかどうかで対応を判断する．

キーワードの使用のみでデータの属性を決定することは，時として抽出結果にエラーを持ち込むことになる．たとえば，多義語（タイ→国名，魚名）の混同や，属性の意味づけの混同（東京→情報発信地，会議開催地，問題対象地）などである．これを避けるには，コンテクストの解析や，注意深いキーワード（属性値）の設定などが，必要になるが，現在のところ，あまり考慮は払われていない．いずれにしろ，多種大量のデータを扱うことを想定する，本システムのようなケースにおいては，あまり有効な解決策は見つかっていない．

3.4 マッパ

データモデルとビューワとの対応を管理するのがマッパである．モデルからビューワへの対応は，各属性について，その属性値を持つ該当データの件数を伝えるのが主な処理である．たとえば，該当データのうち，90年発行のものが31件，91年発行のものが17件，等々という情報を伝えることである．

逆に，ビューワからモデルの属性値（複数も可）を選択して，データを絞り込むプロセスを，メニュー検索と呼んでいる．すなわち，メニュー検索は，任意のモデルにおける属性とその値の組から，それを満足するデータを検索する仕組みであり，DBシステムにおけるスキーマを介した検索と同等の機能である．

その他に，全文検索エンジンを用いた検索が絡んでくることがあるが，これは，図2のパラダイムの外からデータ集合を設定することと考えればよい．

4 SGMLフォーマットによるマルチメディアデータの統一的検索

本システムでは，次の2種類のデータを検索対象とする．白書やJIS規格書などのテキスト情報と，テキストや画像，統計表といった様々な形式の情報を抽象化したメタ情報である．

テキスト情報とメタ情報はSGMLで記述されている．SGMLを採用した理由は，次の点である．

- 文書の論理構造が（データの種別）定義でき，かつ十分な記述力を持っている．これによって，文書構造を保持したままでの可視化が可能となる．また，外部データへの参照も可能であり，文書中の文と図表の関係も定義できる．
- メタデータを，例えば，従来の図書館システムにおける書誌情報のような固定フォーマットでなく，柔軟に記述することが可能である．
- 全文検索において，構造タグを考慮した検索が可能である．

テキストの論理構造は，データの種類によって大きく変化する．本システムでは，3種類のDTD ((1)JIS規格書用 ((財) 日本規格協会制定), (2) その他のテキスト用 (ISO12083), (3) メタデータ用）を用いている．図4に，SGML化された白書データの一部を示す．

本システムでは，データ種別ごとに検索単位が定義されている．例えば，白書の場合，物理的には白書1冊が1ファイルを構成しているが，検索単位は章，節，項であり，検索のためのインデックスを作成する際に論理的に分割される．それぞれの単位の先頭には，対応するSGMLタグがつけられおり，このタグを認識することで，検索単位への論理分割が行なわれる．テキスト以外のデータ，例えば，地域メッシュのような数値情報の場合は，基本的には，一つのファイル（例えば,「1990年の東京都地域メッシュ」）が検索単位となる．

```
<abstract><title>要旨</title>
<section id="sf0.1"><no>第1章</no><title>世界通商・経済の動向</title>
<subsect1 id="sf0.1.1"><no>第1節</no><title>世界経済回復の緩慢さと世界貿易・通商政策</title>
<p>
世界経済は，総体としては緩やかな回復過程にあるものの力強さを欠いており，また，
先進国経済における景気循環局面のずれという短期的側面と，途上国経済の成長や旧共
産圏体制崩壊後の移行という長期的側面が複雑に絡み合った状況にある。世界貿易は，
こうした世界経済の動向を反映して，全体としては92年に伸びを高めた。</p>
</subsect1>
```

図4: SGMLテキストの例

メタデータは，データの種類を問わず，全ての検索単位ごとに定義されている．図5に，メタデータの例を示す．メタデータ中には，対象データ（検索単位）のタイトル，データ種別，出版者等の情報が記述されており，これらが検索の対象となる．テキスト以外のマルチメディアデータは，このメタデータを通じて検索する．

メタデータが検索され，内容表示をユーザに要求された場合，システムの内容表示部は，次のような操作を行なう．

1. メタデータ自体はSGML化されたテキストデータであり，これを可視化する．
2. メタデータが指す実体のデータを判別し，それを可視化できるブラウザを決めて起動する．

本システムでは，あるデータが検索された場合，そのデータに一番適した，あるいはそれを使うユーザに適したブラウザで可視化できるよう設計されている．

例えば，実体データがgif形式の画像データであればxvのような画像ビューワを起動するし，統

```
<meta>
<orgname>
<name>平成五年通商白書＜総論＞</name>
</orgname>
<orginfo>
<orgtype>ＳＧＭＬ</orgtype>
<orgsite>tg93.sgm</orgsite>
</orginfo>
<contents>
<pubbib>
<mtitle>平成五年通商白書＜総論＞</mtitle>
<puborg>大蔵省　印刷局</puborg>
<pubdate>平成５年６月10日</pubdate>
<formtype>Ａ５</formtype>
<npages>419</npages>
<mnum>4-17-270356-9</mnum>
<otherbib>通商産業省</otherbib>
</pubbib>
</contents>
</meta>
```

図 5: メタデータの例

計データであれば，スプレッドシートや3D可視化システムを起動する[1]．

5 情報セットの可視化とナビゲーション

近年，大規模でかつ様々なメディアのデータが電子化され流通しているが，これを情報検索という立場からみると，大規模でかつ様々なメディアのデータからどのように自分の欲しい情報を手に入れるかいう問題がある．根本的にはユーザが欲しているのは検索されたものに含まれる情報であり，これは，検索対象に陽に記述されたものではないことが多い．また，情報相互の関係も，それらを分析してはじめてわかるものあり，これらの情報を可視化し，ユーザに提供する仕組みが必要である．また，一般の検索システムにおいては，ユーザとシステムのインタラクションは，ユーザが欲しい情報を表現した質問（テキストベースの検索システムの場合は単語や文）であり，これが検索対象を適切に表現しているほど，得られる情報も精度が高くなる．しかし，ユーザが，適切な質問を用意できない場合がある．欲する情報が抽象的で具体的な質問に写像できない場合や，そもそも，ユーザが明確な検索意図を持たず発見的に情報を得たい（情報散策）場合である．我々は，情報の可視化やユーザのナビゲーションを統合した機構を Information Outlinging とよび，新聞記事の検索システムを開発している [2]．

以降，本システムでどのようにこれらの仕組みが実現されているかについて説明する．

5.1 検索過程での可視化とナビゲーション

データを様々な角度で検索することを可能にするために，本システムではデータの性質ごとに異なるビューと呼ばれるウインドウを用意している．ビューの機能は次の２つである．

1. ビューに関連するキーワードを提示し，検索を行なう．
2. 検索された情報が，ビュー内のキーワードとどの位関連性が高いかを表示する．

前者の機能は，いわゆる統制キーワードによる検索を行なうための機能である．キーワードを空間的あるいは階層的に表示することで，ユーザの検索行動をナビゲートする．後者の機能は，現在検索されている情報が，どのようなことを表現しているかを把握するために用いられる．ビューは，対象とするデータに合わせて設計される．行政情報を扱う本システムにおけるビューには，次のようなものがある．

- 時系列 (年)
- 日本地図 (県名)
- 世界地図 (国名)
- シソーラス (約２万語の単語の階層関係)
- データソース（例，白書，JIS 規格書，工業会統計）
- データ種別（テキスト情報，メタ情報）

例えば，検索過程でのシソーラスビューの例を図6に示す．上部には，現在表示されているメニューがシソーラス上のどこなのかを示すフィールドがあり，中央部にシソーラス上のあるメニューが表示されている．ユーザは，矢印ボタンを押すことで，シソーラスを渡り歩くことができる．適当なキーワードが見つかれば，それを用いて検索を行なう．

[1] あるデータに対して，それを可視化できるブラウザが複数ある場合には曖昧さが生じるが，現在の実装ではそれを考慮していない．現在は実体のファイル名の拡張子とブラウザを対応させることでブラウザを選択している

ビューの外観も，其々のデータオブジェクトに依存して異なる形式となる．例えば，時系列ビューでは棒グラフ，地図ビューでは，２次元的に配置された地名ボタンの集合，シソーラスでは，階層を上下できるメニュー形式となっている．これに応じて，検索対象との関連度の高さも色の濃さや数値によって表現されている．図7に，日本地図ビューを示す．どちらも，検索結果と県名の関係を示したビューであり，ビューと検索部間の入出力は同じプロトコルである．つまり検索システム側からみれば，どちらも同じビューワクラスのウインドウのインスタンスに過ぎない．データをどのように可視化するかは，ビューワによって決定される．地図ビューを用いると空間的な隣接関係が把握しやすいし，シソーラスビューだと，「日本」の「関東」の「東京」というように，簡単にレベルを上下することができる．

検索が行なわれると，各キーワードの右側に数字が表示される．これが検索された情報と各キーワード間の関連の強さを示している．図6からは，現在絞り込まれているデータが「東京」に深い関係があることがわかる．

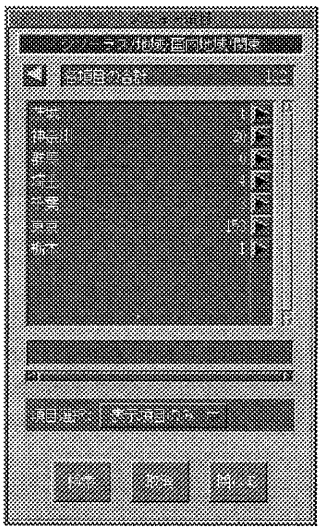

図 6: シソーラスビュー

5.2 検索された情報内容の可視化

前節では，マルチビューを用いてユーザの検索過程をナビゲートする手法について述べた．ユーザは，複数のビューを行き来しながら検索結果を絞りこみ，最終的にその情報の内容を表示させる．

前章で述べたように．本システムでは，データをSGMLで記述しているが，内容表示の際には，検索単位をその都度HTMLに変換し，HTMLブラウザで表示する．HTML変換の仕組みを簡単に説明すると，

1. 検索結果のタイトル表示部から，内容表示を行なう検索単位が選ばれるとHTML変換部に次のような情報を渡す．
   - その検索単位が含まれるファイル名
   - 検索単位のファイル中でのオフセット
   - 検索単位の先頭のタグ
   - 検索式（ブラウザ上でキーワードをマークするため）
   - データ種別

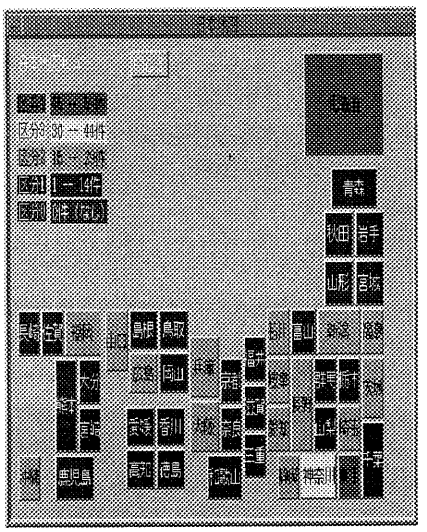

図 7: 日本地図ビュー

2. HTML 変換は，上の情報を受けとると，該当する検索単位をHTMLに変換する．また，「次のデータ」,「前のデータ」,「親のデータ」,「目次」に相当するリンクを自動的に付与する．
3. 変換したHTMLテキストのURLをコントロール部に返し，コントロール部ではHTMLブラウザを起動する．

データの論理的構造を記述するSGMLを用いていることで，検索された検索単位の次のデータへのリンクを計算したり，目次に相当するページを生成することができる．またHTMLブラウザでのデータの表示においても，テキストをフラットな構造でそのまま表示するのでなく，必要に応じて，最初はテキストの骨格のみを表示し，リンクをたどることで，テキスト構造の深部へアクセスすることが可能である．

6 おわりに

新産業創造データベースとして現在準備している主なデータは，

- 白書（通商白書，中小企業白書等）
- JIS規格書
- 日本統計年鑑
- 工業会統計
- 団体調査研究報告書
- 海外産業関連雑誌記事抄録
- 政府資料等書誌情報
- 統計情報インデックス
- 国勢調査
- 事業所統計
- 鉱工業生産指数および第三次産業活動指数

であり，データ容量は，文書・イメージデータが1.7GB，統計データが200MBである．

なお，本報告では，データベースに収容されたデータの検索に関する部分について論じたが，新産業創造データベースセンター自体は，単に，データの検索に留まらず，検索したデータの加工・利用までを視野に入れた実験・実証システムを考えている．実際，新産業創造データベースセンターにより提供されるシステムでは，検索結果としての情報は，クライアント端末の画面で見るだけでなく，クライアント側で稼働しているアプリケーションプログラムに取り込んで，加工することができるような機能を提供している（CSV形式でのデータ提供機能）．

本事業のようにデータベースに蓄えられた情報を，創造的活動に役立てることを目的とするならば，単に情報を探すだけでなく，利用者それぞれが，入手した情報を自らのアイデアで自由に加工することができるような，データ加工の作業場を提供することが重要となるであろう．

ここで紹介したシステムは，IPAが，慶応義塾大学湘南キャンパス内に設けられた情報基盤センターにおいて稼働しており，webを通しても公開されている．本稿では，検索GUIについては，構内で利用するユーザのためのクライアントシステムを中心に紹介したため，多少look & feelが異なるが，基本的に同様の機能が使える．

謝辞

この成果は，情報処理振興事業協会が実施している「新産業創造データベースセンター事業」の一環として行われたものであり，本システムの紹介を許諾いただいた，同協会に感謝いたします．また，実際のシステムの企画・開発に携わった関係各位に感謝します．とくに，SGMLからHTMLへの変換部を担当した凸版印刷株式会社の斉藤伸雄氏，開発に携わった，武田浩一氏，荻野紫穂氏，野美山浩氏，山下晶夫氏をはじめとするIBM東京基礎研究所のメンバーに謝意を表します．

参考文献

[1] 諸橋正幸, 堤泰治郎, 丸山宏, 野美山浩「情報検索システムにおける効果的なナビゲーション機能の提案」, *Workshop on Digital Libraries, No.2*, pp. 45–49, Tsukuba, Nov. 1994.

[2] M. Morohashi, K. Takeda, H. Nomiyama, and H. Maruyama. "Information Outlining - Filing the Gap between Visualization and Navigation in Digital Libraries". *Intl. Symp. on Digital Libraries*, pp. 151–158, Tsukuba, Japan, Aug. 1995.

[3] 小川隆一, 菊地芳秀他「フルテキスト・データベースの技術動向」, 情報処理, Vol.33, No.4, pp.404–412, Apr. 1992.