

情報空間のモデルに基づく情報検索システム

小澤 英昭 中川 透
NTTヒューマンインタフェース研究所

情報検索における情報の高品質化への期待につれて、マルチメディア情報の情報検索が重要となってきた。情報をマルチメディア化することは、同一の情報に多数の表現法を付加することであり、伝達したい情報の周囲の情報が増えて、情報の冗長が大きくなり易い。本稿では、マルチメディア化による情報の冗長を持つデータベースから、効果的に情報を検索するための情報空間のモデル: MVIS (MultiViewpoint Information Space) と、これに基づく情報検索方式を提案する。さらに MVIS に基づく、情報検索システムのプロトタイプについても述べる。

Information Retrieval System
with New Information Space Model for the News

Hideaki OZAWA Toru NAKAGAWA
NTT Human Interface Laboratories
1 - 2356, Take, Yokosuka, Kanagawa, 238-03, Japan

Advanced Information Retrieval Systems are expected now. Some types of the systems are Multimedia Information Retrieval Systems. We will be able to get the information which is represented by the various communication media, such that audio, video, text and images, from the systems. But we might take large unnecessary information from the databases even if we use the traditional information retrieval methods. Because such systems have large amount redundant data in their databases. We propose a new method of the information retrieval for multimedia news databases. It is based on the our proposed new Information Space Model, we call MVIS (MultiViewpoint Information Space). Furthermore we described about the prototype system which is based on the retrieval method.

1 はじめに

大量の情報の中から、必要な情報を見つけ出す情報検索は、計算機を有効に利用する一つの分野として確立されてきた。しかしアクセスできるデータベースが増える情報のマルチソース化につれて、情報は過多になる傾向があり、データベースを検索しても、検索した結果を人間が処理し切れないという、事態に落ちいつている¹⁾。

今後情報のマルチメディア化が進んでいけば、扱う情報は急速に増えるものと考えられ、自分に必要な情報を選択することが、ますます重要になってくる。マルチメディア化とマルチソース化を求める、情報の多様化への要求は、この必要な情報を自ら選択する負荷の増加とのトレードオフの問題に帰着する。

自分に必要な情報をデータベースから検索する手法も、従来のキーワードを使った検索ではなく、マルチメディアやマルチソースの情報に適した検索法を考えることが重要である。

本稿ではマルチメディアの情報検索のための情報空間のモデル MVIS (MultiViewpoint Information Space) と、このモデルに基づく検索手法を提案する。

2 情報検索とマルチメディア

2.1 情報と情報空間

情報検索用のデータベースで扱われる情報には、文献情報であるとか、新聞などのニュース、企業情報、特許情報、特殊なものでは化学式のような情報と非常に多岐に渡っている。しかし扱われている情報の型は、現在の大抵のデータベースでは文字列もしくは数値である。例えばテレビニュースの原稿なども文字列として、情報検索のデータベースで利用できる。

もちろんテレビニュースの場合は、映像情報がついていることがより望ましいが、文字列のみの情報でまがりなりにも情報は、伝達できる。このことは、情報の種類によって最適な表現の仕方は存在するとしても、必ずしも情報を伝達する媒体に依存せずに情報の伝達は行なえるという、情報の本質的な側面を表している。

しかしマルチメディア処理技術が現実の物になるにつれて、情報検索システムに対するマルチメディア化の要求が以下のような点で高まってきている。²⁾

1. 情報の表現法が多彩になる

テキストや数値としてしか表現できなかった情報を、映像や音声を用いて表現できることにより、言語では表現できなかった細かい情報や、雰囲気の情報も表現できる。

2. 情報が理解しやすくなる

絵やグラフを用いることで、直観的に情報を取得しやすくなる。

3. 情報を検索しやすくなる

マップなどを表示することにより、検索要求がだしやすくなる。

これらの要求をまとめると、マルチメディアの持つ表現の技術に期待しているものと考えられる。別な見方をすればマルチメディア化によって、伝達したい情報自身が増加するというよりも、情報検索のシステムと人間の間で情報を伝達しあう際の表現形式の差による点で、情報の増加が生じる。つまりある事件の情報に対して、例えばテキストとして表現された情報に映像の情報が加わったとすれば、現場中継などによる様々の情報が、テキストで表現された時よりも増えてくる(図1)。

本質的には情報が増えていないはずなのに、ユーザが受けとる情報は非常に大きなものとなる結果を生む。結局マルチメディア化された情報の検索では、従来のようにとり合えず網羅的に検索するのでは、情報の冗長が大きいため、受けとる情報の量だけがが増えて、必要な情報が埋没しかねない。

2.2 情報検索の手法

計算機による情報検索を離れて、日常の情報検索の手法について考えてみると、検索と言う動作は、整理もしくは分類と言う動作と対になっている場合が多い。例えば図書館は10進分類法などの分類基準によって、書籍を分類している。また個人的な情報の管理においても、

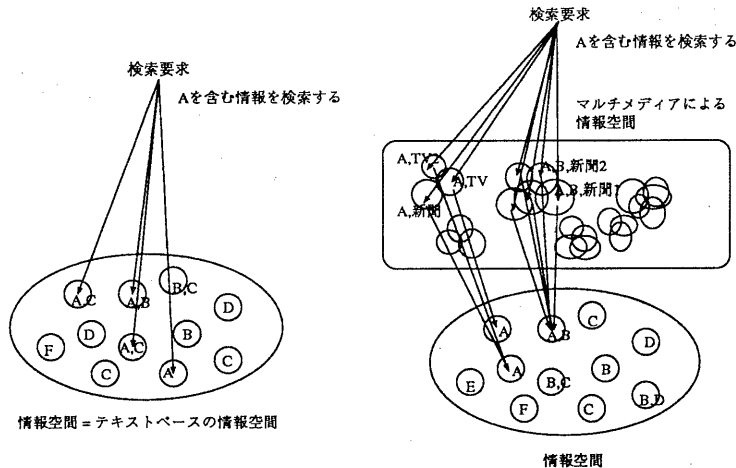


図1. マルチメディアの情報空間

机の上の位置や書棚の位置などの、分類基準によって情報を検索し、常にすべての情報を調べるといったことは行なわない。これは人間の情報検索に対して、次のような仮定を行なっているためである。

- 探せる情報は有限である
全ての情報をチェックして、必要な情報を見つけることはできない。
- 情報間には関連がある
図書館における分類や、本棚の配置、情報が記載されているような書籍の種類など必要とする情報への手がかりがある。

これに対して現在幅広く使われている情報検索のデータベースの検索手法は、キーワードによる検索とフルテキストサーチといった違いは存在するが、基本的に全てのデータを網羅的に調べて条件に該当する結果を全てユーザに与え、ユーザはそこからさらに必要な情報を探す。このような検索法がとられてきた背景には、従来の情報検索のシステムが、元来以下のような仮定を行なっているためと思われる。

- 情報が独立している
情報は情報空間中に無作為に散らばっていて、関連性がはっきりしない。

- 網羅的に情報を必要としている
ある条件を満たす情報を、全てユーザは獲得しようとしている。

前節で述べたように、情報のマルチメディア化が表現形式の差から、類似した情報を多く発生させる結果になると、従来の網羅型の検索では、類似した情報が増える分膨大な量の情報を得てしまう結果になりかねない。そこでマルチメディアの情報に対する、情報検索の概念と手法が必要になると考えられる。

そこでマルチメディアの情報に対する情報検索のサービスは、大きく次の2種類の形態にわかれるものと推測される(図2)。

- 網羅的な検索
マルチメディア化そしてマルチソース化した情報を検索する場合には、網羅的な検索は現在のような単なるキーワード検索ではなく、ある条件に従って本質的に差がある情報のみを検索するサービスとなると推定される。

このタイプの検索は例えば、湾岸危機の全体像を掴みたいような場合に必要となる。表示されるメディアは様々かも知れないが、一連の情報の中で大きく変化した部分だけを選択して提供する。

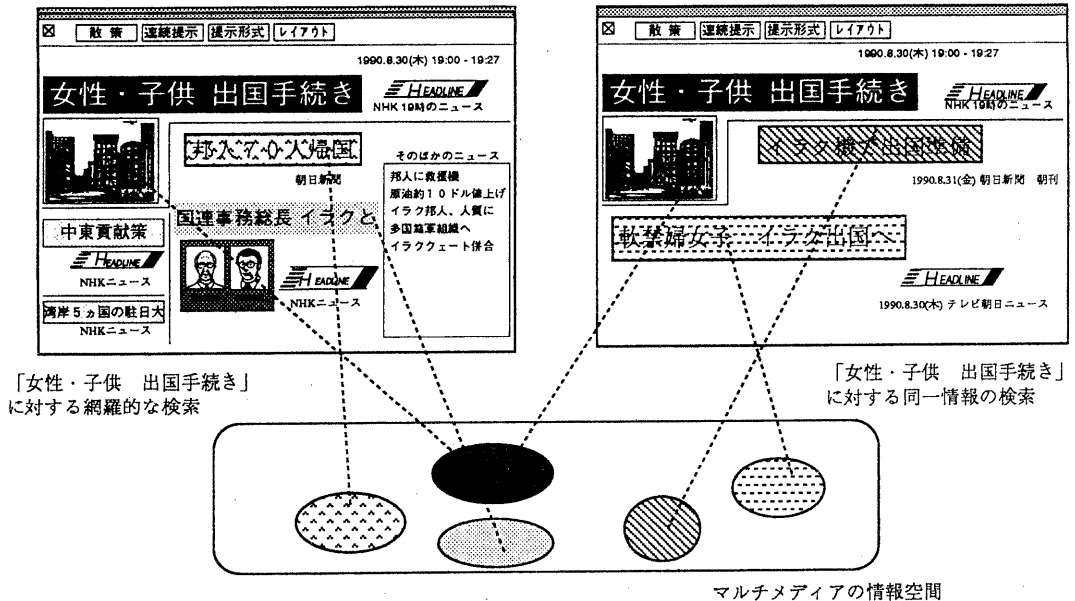


図2. マルチメディア情報の利用法

● 同一情報の検索

網羅的な検索とは逆に、同一の情報と思われる情報について、別のソースによるとらえ方、別の表示メディアによる見え方の違いを見るように、類似した情報のみを提供するサービスが必要となる。

このタイプの検索は、山火事のようなニュースを新聞で知った時に、テレビをつけて現場の雰囲気を知りたいような場合に必要となる。

これらのサービスを実現するためには、従来の情報検索のデータベースのように、情報が空間中に無作為に存在している概念ではなく、日常の探す動作と同じく、情報空間に分類と、関係の強さが表現できるモデルの上に、情報検索システムを構築しなければならない。

3 マルチメディア情報の情報検索

3.1 マルチメディアの情報空間

マルチメディアの情報検索のシステムでは、類似した情報、冗長な情報が増えることに

なるため、情報間の差を定義することが必要である。情報間の差を定義するためには、情報を一定の規則で分類し、ある情報と他の情報とがどれくらい類似しているかの基準を決めることが必要である。これは言い替えると従来の情報検索のシステムで、情報が均一に分布していたとするモデルを、ある情報と属性を基準にして、情報空間中に偏りが生じるかを、決定することにはならない。

そこで本稿では、以下の3つの概念を持つマルチメディア情報空間モデル、MVIS(Multi-Viewpoint Information Space)を提案する。

● 情報の偏りを表現する軸

情報の偏りの軸は、情報を分類法に相当する。図書館のアナロジーで言えば、10進分類法に相当する。情報の世界では、情報の発生した時間、場所等である。

● 情報に関わる範囲

情報間の関係は各軸上で定義されるが、各軸の影響を変化させることで、情報間の関連性の強さも変化する。

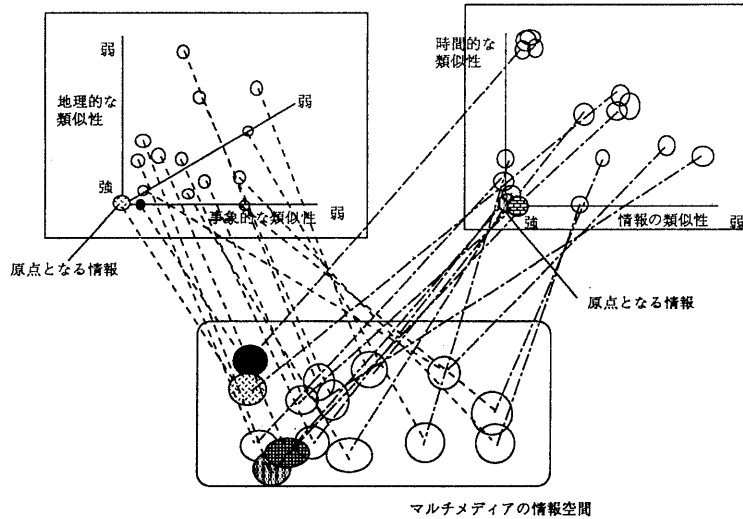


図 3. MVIS による情報空間

● 各軸を束ねる点

情報は絶えず様々の所から発生しているために、発生時間のような軸を別にすれば、情報全体に対して統一的な情報の始点は存在しない。情報間の関連は、ユーザが持っているある情報を基準にし、その情報からの相対的な関連性の強さを扱うものとする。このユーザが最初に示す情報が、軸を束ねる原点となる。

MVIS による情報空間は、図 3 のように表現することができる。MVIS に基づく検索は、従来の網をかける方法から、いわばある軸を作る種をまいて、その種に引き寄せられるもの離れていく物が生じた時に、適当な所を取り出すことによって検索を行なう。軸を束ねる情報は、種になる情報であり、情報空間中のどんな情報でも良い。

情報間の関連性の概念を導入する研究としては、例えば MeSOD³⁾ などがある。しかしマルチメディアの情報は、一つの情報を多面的に利用できるところが重要であり、MVIS では距離的な概念から発展して、情報を空間的な概念として表現した。この 3 つの要素を変化させることによって、情報の空間を様々の視点から見ることができる。このモデルの基づく情報検索

は、即ちユーザが持つ情報を基にして、整列させた空間から、ある部分空間を切り出すことにほかならない。

MVIS に基づく情報検索は、マルチメディア化されたデータベース上で、単に網羅的に情報検索を行なうことによる冗長の排除のために、人間が行なっている検索の手法を、データベースに適用している。

そして提案した検索手法を実現するためには、次の 2 つの技術的課題を、克服することが必要である。

1. 軸の定義と構築

データベースに投入されたデータに対し、情報間の関連の強さを定義する手法を構築する。軸を定義することによって、その一部分空間を定義できる。

2. 原点になる情報の指示法

キーワードの投入に代わり、情報空間中の自分が知っている情報を指示する。

3.2 時事情報データベースの情報空間

ここでは MVIS に基づく情報検索用のデータベースを構築する例として、時事情報のデー

データベースを取り上げる。既に新聞記事やテレビニュースの原稿といったテキストデータを提供するシステムが、多数存在している。新聞社や、テレビ局、ラジオ局といった報道機関は、それぞれの目的に応じて映像や、紙面といった媒体による情報を生成している。これらの情報は同一の情報を各社のとらえ方、並びに各社の固有の情報伝達媒体による情報生成を行っており、複数の報道機関により発せられた時事情報を扱うデータベースは、MVIS に最適なデータベースの例と考えられる。現在の時事情報のデータベースは、キーワードかもしくは掲載された日付を用いて検索する形態をとっており、他の情報検索のデータベースと何ら変わらない。しかし時事情報のデータについて検討すると、次のような特徴が見い出せる。

- 時事情報には偏りが存在する。

時事情報は、人間同士、または人間と自然などとの間における行動の結果として生じる情報であるため、時間的に連続して関連した情報が発生する。

- 複数の表示メディアが存在する。

文字を主とした新聞や、映像と音声为主としたテレビニュースのように、同一の情報を異なった視点と、表現法で伝達するシステムが既に構築されている。

複数の情報の伝達者によって同一の情報が伝達されることも、見方を変えれば情報の偏りと同じであり、時事情報のデータベースは従来の時事情報データベースが仮定してきたような均一の空間ではない、偏りを持った情報空間として考えることができる。

3.3 時事情報と人間の行動

時事情報は、人間と人間、人間と自然、人間と社会との間での、様々の影響や動作を、外から観察した時の見え方を他人に伝える情報として定義する。情報を発生させる原因となっているのは、人間の行動か、人間に対する自然や社会からの影響である。人間の行動は、一見ばらばらで勝手に動いているように見えるが、現実的には様々の制約の上になり立っている。

例えば日本という国であったり、地方であるといった、人間の存在する場所や、戦前という時代や季節といった時間による制約がある。

または時事情報は、例えば PKO 法案の採決のニュースのように、参議院と衆議院で連続して審議が行なわれるといった時間的連続性、時間的因果関係を持っている場合が多い。人間の行なった行動による情報が、次の情報の原因となっていることを示している。

結局時事情報を伝達する事は、これらの行動を起こした人、場所、原因や、行動の対象物、行動の手法という情報を伝達する事に等しい。一般的に時事情報を扱う時には、これらの情報を的確に伝えるために、5W1H(When, Where, Who, What, Why, How) という概念を持っており、新聞記事やテレビニュースは、例えば表 1 に示すように 5W1H に相当する情報を含んでいる。

When	1990 年 9 月 3 日
Where	成田空港に
Who	日本人 70 人
What	
Why	解放されて
How	帰国した

表 1. ニュースにおける 5W1H の例

時事情報のデータでは、例えば 5W1H のそれぞれを、関連性の強弱を決める軸だとすれば、5W1H の各要素の類似度によって、時事情報を MVIS モデル上に表現することができる。

4 MVIS モデルに基づく時事情報検索システム

4.1 軸の構成法

実装したシステムでは、時事情報が持つ言語化された情報から、5W1H を表現する言葉の抽出を自然言語処理の技術を用いて行なった。処理に用いたデータは新聞では本文、テレビニュースではスクリプトである。

具体的な処理法としては、時事情報を形態素解析し、品詞と 5W1H の関係を調査し、その結果単語のレベルで見た場合には、5W1H による情報の表現の軸は、大き

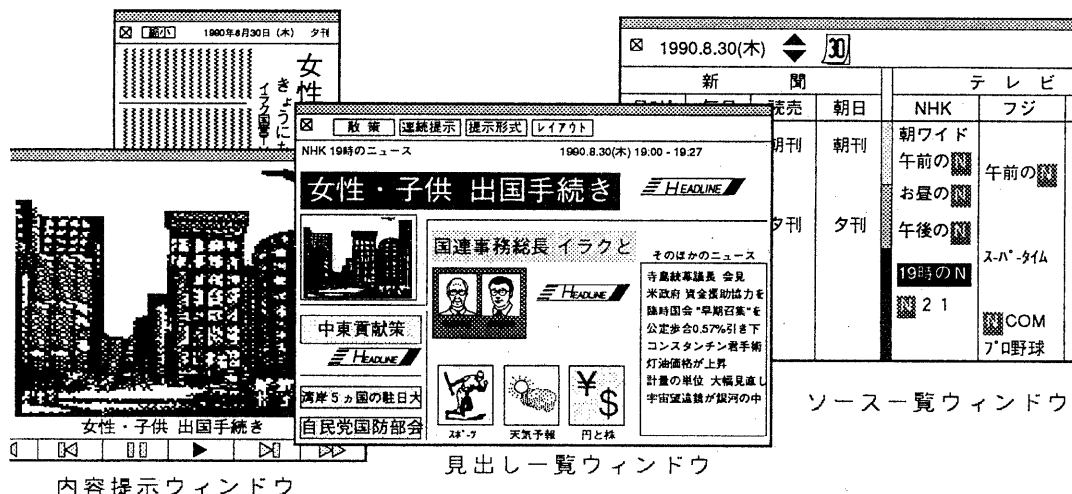


図 4. 情報散策システムによる原点情報の獲得

く Where, Who, What と言った情報が発生した対象や、場所と言った情報を生じさせるものに依存した軸と、How, Why といった動作自身の情報の大きく 2 つに分類することができる。この 2 種類の情報は、主に物を表現する固有名詞と、動作を表現するサ変名詞 (～するという言葉) によって分離される。そこでシステムでは、情報に関わったものに対する軸と、情報自身の動作による軸を持つ空間として表現する方法を採用した。

この外に時事情報には、名詞が数多く存在し、名詞によって個々の情報は大体分離される。また既に述べたように、時事情報は因果関係に従って時間軸に沿って情報が発生することから、時事情報中の時間に関する情報から、時間の軸を作成し、内部的には都合 4 軸の情報空間として表現して、関連度計算のアルゴリズム⁴⁾を用いて、関連性の強さの計算している。

4.2 情報の検索法

4.2.1 原点になる情報の獲得

MVIS に基づく情報検索システムでは、ユーザからの検索要求として、原点となる情報

の入力が必要である。MVIS に基づくシステムを有効に利用するためには、原点になる情報は、従来のキーワードの投入に比べて、多くの情報が必要である。そこで本システムでは、ユーザが直接情報を入力する代わりに、ユーザが類似していると考える情報をデータベースから選択する機構を用い、ユーザが選択したある情報を原点の情報とみなして MVIS に基づいて情報検索を行なう。

このために本システムでは、図 4. に示す情報散策システムのニュースソース一覧ウィンドウと見出し一覧ウィンドウを、最初の情報を獲得する際のインタフェースとしている。⁵⁾

4.2.2 検索範囲の設定

MVIS に基づく検索システムでは、情報の偏りを表現する軸の影響の強さと、範囲を指定することで、検索する情報を制御する。作成したプロトタイプシステムでは、これらの設定を行なう例として、検索パレットを提供している。検索パレットは、システム内部の 4 軸の内、時事情報の対象を示す固有名詞の軸と、事件の内容を表現するサ変名詞の軸を制御する、「関係の方向」と呼ぶスライドバーと、時間軸

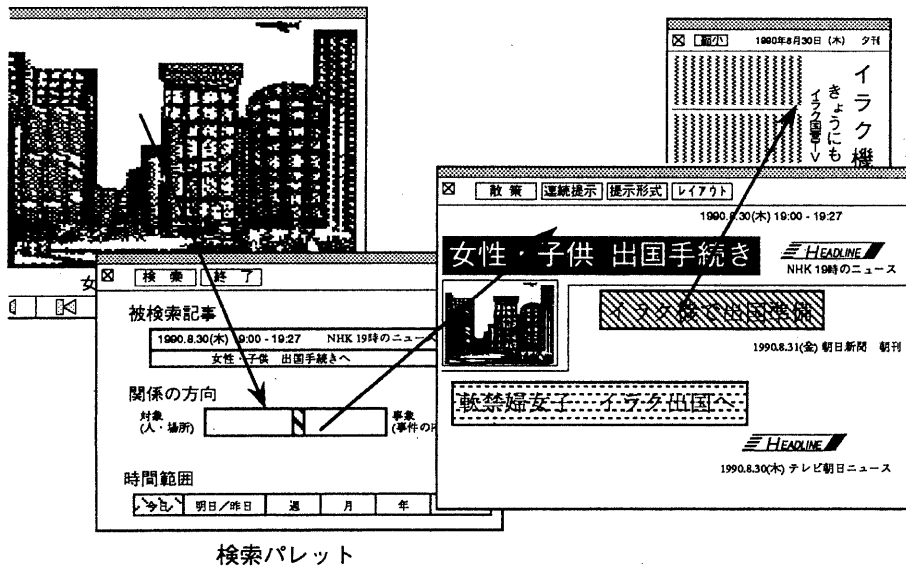


図5. メディアの違いによる同一情報の検索

に対して検索したい範囲を指示する「時間範囲」を持っている。

これらの設定を変化させることにより、例えば「時間範囲」で「今日」を選択し、「関係の方向」で、「対象、事象」の中間を選べば、原点となる情報に類似した当日の情報、即ち他の報道機関が報道した情報を獲得できる。

5 まとめ

本稿では情報検索のデータベースがマルチメディア化された状況を検討し、マルチメディア情報のデータベースの検索手法を提案した。

この手法は、無闇にデータベース全体を検索するのではなく、情報空間に偏りを持たせたモデルと、それに基づいて部分空間を抽出する手法である。

本手法は、情報がマルチメディア化されることにともなって、表現の違いによる多くの情報が提供され、キーワードなどを用いた従来の情報検索では、情報が過多になってしまう状況において、有効な手法と考えられる。本稿では時事情報のデータベースを例として、提案したモデルにおける情報検索の手法を説明した。

今後の課題としては、時事情報においては

5W1Hのような、人間の行動を表現する情報が、偏りの軸の一つとなったが、他の情報の偏りの表現の軸を探し、体系づけていくことが必要である。

参考文献

- 1) L. F. Rau and P. S. Jacobs, "Creating Segmented Databases From Free Text for Text Retrieval", Proc. of 14th Annual International ACM SIGIR Conference, PP 337-346, Oct. 1991
- 2) 三輪 真木子, "情報検索システムにおけるユーザインタフェースの条件", 情処研報 Vol. 91 No. 77, 91-IM-3-3, Sep. 1991
- 3) 田淵 仁浩, 村岡 洋一, "距離空間データモデル MeSOD におけるオブジェクトの概念について", 信学論 D-I 分冊, Vol. J74-D-I No.4, pp296-306, Apr. 1991
- 4) 小澤 英昭, 中川 透, "マルチメディアデータ間での言語情報を用いた関連づけの手法", 情処第44回全国大会 4-147, 1992
- 5) 大久保 雅且, 小林 直樹, 高田 久靖, "ダイナミックペーパー型対話通信システム(2) - 情報散策方式の検討, 情処第44回全国大会 3-363, 1992