

論文 / 著書情報
Article / Book Information

題目(和文)	超分散環境における情報探索に関する研究
Title(English)	
著者(和文)	菅原真司
Author(English)	
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第4019号, 授与年月日:1999年3月26日, 学位の種別:課程博士, 審査員:
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第4019号, Conferred date:1999/3/26, Degree Type:Course doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

超分散環境における
情報探索に関する研究

[学位論文]

1999 年 1 月

指導教官 酒井 善則 教授

提出者 大学院 理工学研究科
電気・電子工学専攻
菅原 真司

目次

1 序論	1
2 超分散データベースとその探索問題	3
2.1 超分散データベース	3
2.2 探索問題	4
2.3 探索情報の存在確率	5
3 エントロピーを用いた探索手法	8
3.1 まえがき	8
3.2 情報探索問題の定式化	8
3.3 情報探索問題とエントロピー	8
3.4 具体例による検証	9
3.5 画像探索への応用例	11
3.6 まとめ	13
4 効率的画像情報探索システム	14
4.1 まえがき	14
4.2 超分散データベースにおける画像情報探索	14
4.2.1 情報探索の効果	14
4.2.2 EBS 法	15
4.2.3 各ノードにおける被探索情報の存在確率	16
4.3 画像情報探索方式	17
4.3.1 画像情報の探索手法	17
4.3.2 探索手法の仮定	18
4.3.3 探索手法の組み合わせ	19
4.3.4 探索アルゴリズム	20

4.3.5	解析	20
4.4	色対情報を用いた類似画像検索手法	22
4.4.1	色対による検索方式	23
4.5	実験システム	23
4.5.1	システムの動作と探索対象情報	23
4.5.2	システムの構成	23
4.5.3	実験結果	26
4.6	まとめ	27
5	画像情報の効率的探索手法	29
5.1	まえがき	29
5.2	超分散データベースとその探索問題	29
5.2.1	超分散環境に散在する情報源	29
5.2.2	探索問題	30
5.3	探索の戦略	30
5.3.1	探索方式	31
5.3.2	最適性の検証	31
5.4	特性解析	33
5.4.1	提案方式の特性解析	33
5.4.2	ターゲット存在確率の誤差感度	36
5.5	方式の簡略化	37
5.6	まとめ	39
6	エージェントを用いた探索	41
6.1	まえがき	41
6.2	エージェントの概要	41
6.2.1	エージェント型探索の特長	41
6.2.2	エージェントの定義	42
6.3	情報探索問題	43
6.3.1	通信・処理コスト	43
6.3.2	時間コスト	43
6.3.3	ノード間の回線容量	44
6.3.4	各ノードにターゲットの存在する確率	44
6.4	エージェントモデルとその分類	45

6.5	Single-Fixed Type モデル (基本モデル)	46
6.5.1	探索問題	46
6.5.2	最適探索法	48
6.6	Single-Mobile Type モデル	50
6.6.1	探索問題	50
6.6.2	探索アルゴリズム	51
6.6.3	提案アルゴリズムの評価	52
6.7	Multi-Fixed Type モデル	55
6.7.1	探索問題	55
6.7.2	提案アルゴリズム	56
6.7.3	アルゴリズムの計算量	58
6.7.4	探索アルゴリズムの評価	59
6.8	まとめ	61
7	探索打ち切りの効果	62
7.1	まえがき	62
7.2	インターネットにおける情報探索	62
7.3	統計的手法を用いた探索打ち切り	64
7.4	探索報酬とペナルティを考慮した探索打ち切り	65
7.4.1	探索報酬とペナルティ	65
7.4.2	探索手法	65
7.4.3	数値解析	66
7.5	まとめ	71
8	結論	72
	参考文献	75
A	5.3 節の定理の証明	78
B	$E(n_j + n_k) > E(n_k)$ の証明	82
C	$E(n_1 + n_2 + \dots + n_n) > E(n_n)$ の証明	83

第 1 章

序論

昨今の技術革新に伴い、計算機ハードウェアおよびソフトウェアの性能は、驚くべき速さで飛躍的な進歩を遂げつつある。これにより、非常に高性能かつ低価格のパーソナルコンピュータが、誰にでも簡単に手に入れられるようになった。同時に、ISDN やインターネットに代表される大規模ネットワークが整備され、これに自分のコンピュータを接続するユーザの数は増加の一途をたどっている。

個人ユーザが扱うことができるデータの種類や容量も以前では考えられないほど多種、多様、大容量となり、ユーザ間でのこれらの情報の交換はもはや日常的であると言って良い。

大規模ネットワークにおいて、それに接続する各個人ユーザがそれぞれ自分の持つ情報を、他の不特定のユーザに対して自由に閲覧できるように提供した場合、そのネットワーク全体は、各個人ユーザから見て、データベース的な機能を持つ巨大な情報空間と考えられる。しかし、通常のデータベースとは異なる特性を持つため、ユーザが必要とする情報をネットワークから効率的に取得するためにはいくつかの問題点もある。

このような見方はインターネットの WWW(World Wide Web) においては既に現実のものとなり、情報探索のためのブラウザ、エージェント等が研究・開発されている [1]~[16]。

しかし、これらの研究・開発ではネットワーク内の情報収集を目的としたシステムの機能分担に重点がおかれ、情報検索を認知的観点で扱ったような研究は見られるものの [17][18]、目標を発見するまでの効率については理論的な研究はなされていない。一方、情報検索の効率を対象とした研究は、文書検索を中心に多くの著述が見られるが、ネットワークにおける探索効率という観点での研究は殆どない。

本研究では、ネットワーク内に独立した情報源（ノード）が複数存在する状態を超分散データベースと呼び、このようなデータベース内から目的とする情報（ターゲット）を効率良く探索する手法について、検討を行うものである。

超分散データベース探索問題は、ネットワーク内に存在する、ターゲットを所有するノー

ドの個数により、問題が異なる。本研究では、ネットワーク内における、ターゲットを所有するノード数が1である場合を単一標的問題、不定（単数、複数の場合、あるいは存在しない場合も含む）である場合を不定標的問題として、これら2つの問題について論じている。

また、どちらの場合もネットワーク内の各ノードにおける、ターゲットの存在確率は既知であると仮定するが、実際には様々な条件からユーザが推定できるものと考えられる。これには多くの方法が考えられるが、本研究では、各ノードにおけるターゲットの存在確率推定手法の1方式について触れ、その推定された確率の誤差の影響についても考察している。

特に不定標的問題に関しては、エージェントを用いた情報探索手法と、探索打ち切りの効果についても論じた。前者については、実際に探索を行うエージェントの数とそれらの機動性の有無に着目して分類し、そのうちのいくつかの方式について、効率的な探索手法を提案し、評価を行った。また、後者については、大規模なネットワーク内のノードを全て探索することは事実上不可能であるため、実際にはある時点で探索を打ち切る必要があるが、このための明確な指針を示すための1方式について提案している。

以下の第2章では、超分散データベースについて述べ、その探索問題について、単一標的問題と不定標的問題の2つを定義する。この後、第3章と第4章では単一標的問題に、また、第5章から第7章までは不定標的問題に、それぞれ分類される事柄について述べる。まず、第3章では、単一標的問題の1解法としてのエントロピーを用いた手法について述べる。また、第4章では第3章で述べた手法を用いて、画像情報探索システムの提案、試作を行っている。第5章では、画像情報の効率的探索手法について論じ、ターゲットの存在確率の推定誤差の感度に関しても考察する。第6章では、エージェントを用いた情報探索手法について述べ、第7章では探索打ち切りの指針について示している。最後に第8章で、本研究の結論を述べる。

第 2 章

超分散データベースとその探索問題

2.1 超分散データベース

超分散データベースモデルを図 2.1 に示す。

ISDN やインターネットに代表される大規模なネットワークにおいて、多くのユーザが個人の計算機資源を接続し、互いに情報の交換を行っている環境を考える。ここで、各ユーザが特定の相手との情報の送受信を行うだけではなく、不特定多数のユーザに対して自分の持つ情報を発信する場合、あるいはそのような情報発信を行っている不特定多数のユーザから自分が必要とする情報を受信する場合、あるひとりのユーザから見て、この大規模ネットワークはひとつの巨大なデータベースと考えることが可能である。

現実には、インターネットにおける WWW(World Wide Web) では、個人や団体が自由に WWW サーバを立ち上げ、インターネットに接続することで、世界中のインターネットユーザに対して、情報を発信することが可能である。同時に他の同様のサーバにアクセスすることにより、自由に様々な情報を受信することも可能である。

このようなデータベースでは、各ユーザは大規模通信ネットワーク上で、他のユーザが提供する情報の中から必要とするものを自由に探索し、入手することが可能である。

超分散データベースモデルは、従来のデータベースと比較して以下の 2 点が特徴となる。

- 情報の内容およびその存在位置については、一元的な管理は行われない。
- 提供される情報の所有者は、それを提供するユーザ自身であり、その意思により常に自由な更新、削除が可能である。

よって、超分散データベースは、通常のデータベースよりも柔軟かつ穏やかに統合されたネットワーク上の情報の集合であると言える。

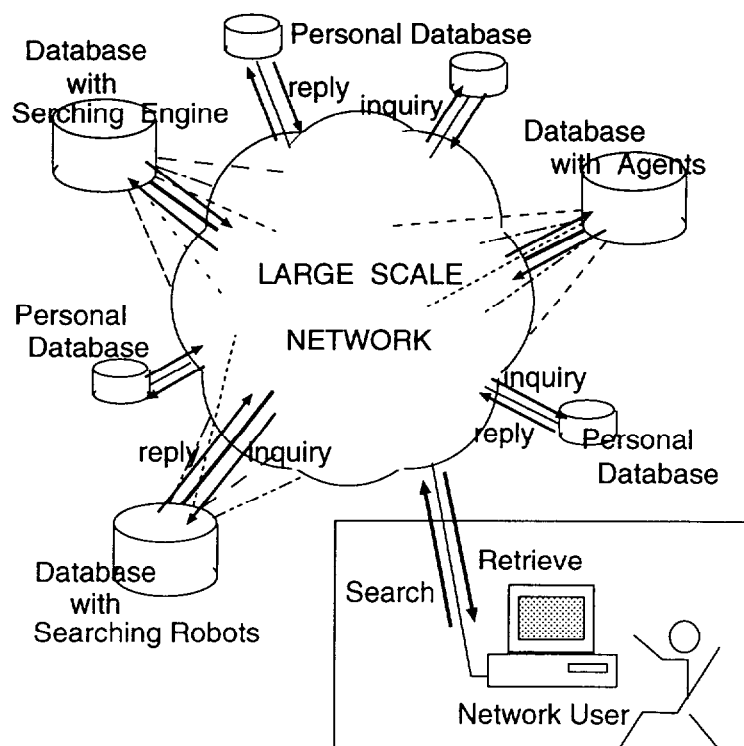


図 2.1: 超分散データベースモデル

2.2 探索問題

超分散データベースを構成する要素として、多数の情報源および情報通信用資源（ノード）とそれらを接続する大規模通信ネットワークを仮定する。探索者はこのネットワークに接続された1ノードから他のノードに対して目的とする情報（ターゲット）を求めて探索を行うものとする。

このようなデータベースにおいてターゲットを探索する場合、無作為な探索は、非常に効率が悪いことは明らかである。よって、本研究では超分散データベースの探索問題について、

「探索にコストを考慮し、効率的な探索を行う手法を求める」

と定義する。

本研究では、コストとしてまず第一に、探索に要する通信・処理コストを考える。これ

は、探索者から、探索対象となる各ノードまでの通信コスト、および探索先ノードに存在する情報がターゲットであるか否かの判断を行うための計算処理コストからなる。さらに問題によっては、通信および計算処理に要する時間も、利用する回線や計算資源により異なるため、探索が終了するまでに要する時間もコストとして考える。

この問題を検討するにあたっては、次の2つのモデルを対象とする。

[モデル 1]: ターゲットの存在するノード数はネットワーク全体に1箇所（単一標的）

[モデル 2]: ターゲットの存在するノード数は不定（不定標的）

これらの2つのモデルはそれぞれ2つの違った探索問題として扱うことができる。

2.3 探索情報の存在確率

本研究では、各ノードにおけるターゲットの存在確率を用いた議論を行う。

各ノードにおけるターゲットの存在確率は、上で述べたモデル1と2で定義が異なり、前者の場合には、ターゲットの存在確率の総和が常に1であるのに対して、後者では、存在確率が各ノードにおいて独立であると考えられる。

<モデル 1 : 単一標的問題>

モデル1の場合は、各ノード間のターゲット存在の確からしさの比が推定できると考える。また、ターゲットが存在するノードがネットワーク全体に1箇所であるから、各ノードにおけるターゲット存在の確からしさの総和は1となる。1回の探索（一般には複数のノードを同時に探索）を行った結果、ターゲットが発見できなかった場合、その際に探索したノード集合にはターゲットが存在しないことが判明する。これは、ターゲットの存在するノードが、未探索のノードに限定されることになり、結果として未探索ノードにおけるターゲット存在の確からしさが上昇することになる。モデル1では常にこのような確からしさの正規化（総和を1にすること）を行うことができる。

従って、本研究のモデル1では、あるノード i におけるターゲットの存在確率 P_i は、ノード総数を N として、

$$\sum_{i=1}^N P_i = 1 \quad (2.1)$$

また、各ノード i における探索前のターゲットの存在確率を P_i 、ある1回の探索によって探索されたノードの集合を S 、探索後、ターゲットが発見されなかった場合の、各ノード

の存在確率を P'_i として,

$$P'_i = \frac{1}{\sum_{k \notin S} P_k} P_i \quad (i \notin S) \quad (2.2)$$

$$P'_i = 0 \quad (i \in S) \quad (2.3)$$

と書けるものとして定義する.

このようなモデルの例としては, 複数のノードからなる画像データを管理するデータベースにおいて, ターゲット画像がどこか 1 箇所のノードに必ず存在する場合を想定することができる. この画像データを画像特徴量マッチングを用いて探索する場合, 特徴量ベクトルの一致度とターゲット存在の確からしさには相関があると考えられる. 従って, 未探索のあるノードと他のノードとの一致度の比が求められるため, 常に未探索ノード同士のターゲットの存在確率の比は一定であると推定される.

<モデル 2 : 不定標的問題>

モデル 2 の場合を説明するにあたって, あるターゲットが存在するか否かということについて, 全く同様に確からしいと考えられるノードが, 仮に N_s 箇所存在する場合を考えよう. ここで, 実際にはターゲットは N_s 箇所中 $n_s (< N_s)$ 箇所のノードにのみ存在するとき, これ以上の情報がなければ, 先の N_s 箇所のどのノードにも等しく n_s/N_s の確率でターゲットが存在すると考えられる.

逆にこのような確率 P_i が全てのノードに与えられたとき, ターゲットが等しい確率 $P_i = P_0$ で存在する複数のノードを n 箇所探索した場合, そのうち $P_0 \cdot n$ 箇所からターゲットを発見することができる.

モデル 2 では, この確率 P_i をノードにおけるターゲットの存在確率と定義する.

このようなモデルの例としては, インターネット上において情報探索を行う場合に, 情報検索エンジンにより提供される, 探索対象とすべきサイト情報をもとにした探索が該当する. 多くの情報検索エンジンでは何らかの方法で, ユーザの指定したターゲットが存在すると推定されるサイトおよびそのターゲットの存在の確からしさを提示する. これらのサイトを探索対象とし, 提示されたターゲット存在の確からしさを存在確率として用いた場合, モデル 2 と同様のモデルとなる.

モデル 1, 2 のいずれの場合でも, 確率の推定には誤差を含むため, 誤差感度はあまり大きくない手法を用いる必要がある.

これらの2つのモデルについて、これ以降、モデル1の単一標的の問題については第3章と第4章、モデル2の不定標的の問題については第5章から第7章でそれぞれ述べる。

第 3 章

エントロピーを用いた探索手法

3.1 まえがき

本章では、超分散データベースの単一標的問題において、情報探索を最も効率良く行うための手法の理論的検討を行う。具体的には情報の存在場所の曖昧さがエントロピーに対応することに注目して、エントロピーの減少が最大となる探索法が最も望ましい探索法であると考え、この方式の有効性を示すと共に、画像情報探索への適用法を提案している。

3.2 情報探索問題の定式化

情報探索問題を以下のように定式化する。

n 個の情報源 DB_i ($i = 1 \sim n$) があり、その一つに目的の情報が存在する。 DB_i に情報が存在する確率を p_i とする。情報探索の行為を ACTION (A) と呼ぶ。ACTION にはコストが付随する。ACTION A_j に対するコストを C_j とする。 A_j により $p_i \rightarrow p_i'$ と変化する。情報探索問題は同一コストの ACTION の中で最も情報探索の効率の良い方法を選択することにある。

一般に情報探索問題では、平均探索回数あるいは最大探索回数を最小にすることが望ましい。このためには情報を発見するまでの手順が必要となるが、本章では1回の探索にのみ着目して、その時点で望ましい探索法を選択できるための基準について考察する。

3.3 情報探索問題とエントロピー

情報が存在する世界の中で情報源のある場所の不確かさはエントロピー H

$$H = - \sum_{i=1}^n p_i \log p_i \quad (3.1)$$

で表される [19]. 例えば, 情報源の存在する場所が既知である場合は, $p_m = 1$ である m が 1 つ存在し, $H = 0$ となる. 探索は H を H' に減少させる効果がある. 従ってここでは,

$$\Delta H = H - H' \quad (3.2)$$

を探索の効果と考え, ΔH 最大の探索が最も望ましいと考える. なお, 以後エントロピーに基づいた探索法を EBS(Entropy Based Search) 法と呼ぶ.

3.4 具体例による検証

ここでは典型的な具体例をもとに 3.3 で述べた EBS 法の正しさを示す.

(1) 順次探索

$DB_i (i = 1 \sim n)$ を毎回 1 つずつ探索する方式を考える. このとき, DB_i の探索順が重要な選択になる.

DB_m を選択した場合,

- 選択が正しければ, エントロピー $H_1 = 0$.
- 選択が正しくなければ, 残りの $DB_i (i \neq m)$ での情報の存在確率 p_i は, 次のようになる.

$$p_i' = \frac{1}{\sum_{k \neq m, k=1}^n p_k} p_i \quad (3.3)$$

従って, エントロピー H_2 は,

$$H_2 = - \sum_{i \neq m, i=1}^n p_i' \log p_i' \quad (3.4)$$

以上より, ACTION 後のエントロピーの期待値は,

$$\begin{aligned} H' &= p_m H_1 + (1 - p_m) H_2 \\ &= -(1 - p_m) \sum_{i \neq m, i=1}^n \frac{p_i}{\sum_{k \neq m, k=1}^n p_k} \log \frac{p_i}{\sum_{k \neq m, k=1}^n p_k} \end{aligned} \quad (3.5)$$

となる.

$$\sum_{k \neq m, k=1}^n p_k = 1 - p_m \quad (3.6)$$

より,

$$H' = - \sum_{i \neq m, i=1}^{i=1} p_i (\log p_i - \log(1 - p_m)) \quad (3.7)$$

従って, エントロピーの減少量 ΔH は,

$$\begin{aligned} \Delta H &= H - H' \\ &= -p_m \log p_m - (1 - p_m) \log(1 - p_m) \end{aligned} \quad (3.8)$$

となる. このことから次の性質が成り立つ.

「 ΔH を最大にする p_m は p_i の中の最大値となる」

(証明)

$H = f(p_m)$ と考えると, $f(p_m)$ は $0 \sim 0.5$ で単調増加, $0.5 \sim 1$ で単調減少, かつ 0.5 の前後で対称となる関数となる. p_i の最大値を p_{max} とする.

(i) $p_{max} \leq 0.5$ の場合

全ての p_i が 0.5 以下であり, $f(p_m)$ が単調増加の範囲であるから, $f(p_m)$ を最大とする p_m は p_{max} となる.

(ii) $p_{max} \geq 0.5$ の場合

$$\sum_{i=1}^n p_i = 1 \text{ より,}$$

$$p_{max} - 0.5 = 0.5 - \sum_{i \neq max, i=1}^n p_i \quad (3.9)$$

となり, これより

$$p_{max} - 0.5 \leq 0.5 - p_i \quad (i \neq max) \quad (3.10)$$

となる. $f(p_m)$ は $p_m = 0.5$ で対称であるから, $p_m = p_{max}$ が $f(p_m)$ を最大にする.

(証明終)

一方, DB_i を探索順に並べて \hat{DB}_i と記した場合, 情報を発見するまでに必要な平均探索回数は $\sum_{i=1}^n i \hat{p}_i$ となる. 明らかに,

$$\hat{p}_1 \geq \hat{p}_2 \geq \hat{p}_3 \geq \dots \quad (3.11)$$

とした場合、平均探索回数は最小となるが、これは結果的に EBS 法と一致する。

(2) 2分木探索

DB_i を常に2つのグループに分けて、1回の探索でどちらのグループに情報があるか探索する方法を考える。この場合 DB_i のグループ分けの方法の選択が重要となる。簡単のため、 DB_i は p_i の昇順に並んでいるものとする。 DB_i を $DB_j^1(j=1\sim n_1)$, $DB_j^2(j=n_1+1\sim n)$ に分けるものとする。

1回の ACTION 後のエントロピーは、

$$\hat{p}_1 = \sum_{j=1}^{n_1} p_j \quad (3.12)$$

$$\hat{p}_2 = \sum_{j=n_1+1}^n p_j \quad (3.13)$$

とすると、

$$\begin{aligned} H' &= \hat{p}_1 \left[- \sum_{j=1}^{n_1} \frac{p_j}{\hat{p}_1} \log \frac{p_j}{\hat{p}_1} \right] \\ &\quad + \hat{p}_2 \left[- \sum_{j=n_1+1}^n \frac{p_j}{\hat{p}_2} \log \frac{p_j}{\hat{p}_2} \right] \\ &= \left[- \sum_{j=1}^{n_1} p_j \log p_j - \sum_{j=n_1+1}^n p_j \log p_j \right] \\ &\quad + \sum_{j=1}^{n_1} p_j \log \hat{p}_1 + \sum_{j=n_1+1}^n p_j \log \hat{p}_2 \end{aligned} \quad (3.14)$$

となる。従って、

$$\Delta H = -\hat{p}_1 \log \hat{p}_1 - \hat{p}_2 \log \hat{p}_2 \quad (3.15)$$

であり、これを最大にするためには、 \hat{p}_1 , \hat{p}_2 をできる限り $1/2$ に近づければ良い。2分木探索法では常に2つの木の確率が等しくなるようにすることが、平均探索回数を最小とすることが知られている。従って、EBS 法は結果的に最適手法となっている。

3.5 画像探索への応用例

画像を探索する場合の重要な方法として、キーワードによる探索とブラウジングがある。キーワード探索は、探索手続きは容易であるが完全に正しい画像を発見することは困難で

ある。キーワード+ブラウジングにより完全な探索が可能となるが、ブラウジングには大きな探索コストがかかる。ここでは次の仮定をおき、問題を考える。

(1) n 個の DB_i があり、そのうち 1 箇所探索対象画像が存在し、存在確率 p_i は既知である。

(2) キーワード探索により、各 DB_i から 0, 1 の回答が返信されるが、その確率は次の通りである。

- DB_i に探索対象画像が存在する場合

1 が返信される確率 1

0 が返信される確率 0

- DB_i に探索対象画像が存在しない場合

1 が返信される確率 \hat{p}

0 が返信される確率 $1 - \hat{p}$

(3) キーワード+ブラウジングを行うと、(2) において $\hat{p} = 0$ 、すなわち完全な探索が可能となる。

(4) キーワード+ブラウジングの合計探索コストはキーワードの m 倍である。

ここでは簡単のため、 $n = m$ 、すなわち全 DB に同時にキーワード探索することと、1 DB にキーワード+ブラウジング探索をすることが同一コストとして、次の問題を EBS 法により考察する。

「 p_i, \hat{p} が既知の場合、次の ACTION のうち、どちらを選択するか」

ACTION 1:

全 DB_i に同時にキーワードを送信し、1 を返信した DB_i のみ以後の探索対象とする。

ACTION 2:

$p_m = \max p_i$ の DB_i に対してキーワード+ブラウジングの探索を行う。

ACTION 1 で 1 を返信した DB_i に対しては、以後ブラウジングにより順に探索を行うこととして、ACTION 2 で探索が失敗した場合は、残りの DB_i に対して同様な手続きを繰り返す。明らかに ACTION 1, ACTION 2 の探索コストは同一であり、ここでは EBS 法による ACTION の選択を試みる。なお、EBS 法はあくまでもその ACTION 時点でのエントロピー減少量のみを考慮しており、以後の探索でどちらが有利かは完全には解らない。ACTION 1 を行った場合、1 を返信する DB_i の数を考える。探索対象画像が存在する DB_i

からは必ず1が返信される。従って、1を返信する数が k となる確率 $P_r(k)$ は次のようになる。

$$P_r(k) = {}_{n-1}C_{k-1} \hat{p}^{k-1} (1 - \hat{p})^{n-k} \quad (3.16)$$

k 個の DB_i における探索対象画像の存在確率は、事前存在確率 p_i と、キーワード探索で1が返信されたことによる情報で定まるが、ここでは簡単のため、1を返信した DB_i での探索対象画像の存在確率は全て等しく $1/k$ と考えると、ACTION後のエントロピー H' は、

$$H' = - \sum_{k=1}^n P_r(k) \frac{1}{k} \log \frac{1}{k} \times k \quad (3.17)$$

従って、エントロピーの減少量は、

$$\Delta H_1 = - \sum_{i=1}^n p_i \log p_i + \sum_{k=1}^n P_r(k) \log \frac{1}{k} \quad (3.18)$$

となる。これに対して、ACTION 2を行った場合、

$$\Delta H_2 = -p_m \log p_m - (1 - p_m) \log(1 - p_m) \quad (3.19)$$

となる。従って、 ΔH_1 と ΔH_2 の比較より選択が可能となる。

n と m が一致しない場合は、1回の探索ACTIONコストを等しくすることが困難となるので、(エントロピーの減少量)/(探索コスト)等をACTIONの判断基準にすることも考えられるが、いずれにせよエントロピーの減少量が探索の効果を表すことは明らかである。

3.6 まとめ

本章では、情報探索を効率良く行うために、エントロピーをもとに探索方法を選択する方式を提案して、その有効性を示した。さらに、画像探索における本手法の適用可能性について論じた。

第 4 章

効率的画像情報探索システム

4.1 まえがき

画像情報を探索する場合、従来は言語系情報であるキーワード、直接目で見えるブラウジングが用いられ、WWW でもほとんどこれらの手法が用いられているが、最近では類似画像検索のように画像情報自体を検索キーとして用いる試みがなされている。

とりわけネットワーク内の画像情報探索にあたっては、どの探索方法で、どのデータベースから情報探索を行うかが極めて重要になる。

本章では特にエントロピーを尺度とした情報探索方式の画像探索への適用法の提案、自然画像の高速検索に適した色対検索法の提案、およびこれらの方式を用いた画像情報探索システムの試作結果について述べている。

以下 4.2 では、3 節で述べた EBS 法を実際の探索に用いるための準備を行い、4.3 ではその画像探索方式について述べた。4.4 では画像探索に用いる手法として特に色情報に着目した類似画像探索手法について触れ、4.5 では以上の検討に基づいた、画像探索プロトタイプシステムの設計・製作と実験結果について述べている。

4.2 超分散データベースにおける画像情報探索

ここではエントロピーをもとにした探索方式について述べる。

4.2.1 情報探索の効果

画像情報探索では後述するように様々な探索方式の併用となるため探索にかかる総コストが問題となる。このとき、各ユーザが提供するデータベース内の情報源 (以下では、それ

らをノードと呼ぶ) に対する探索順の決定は重要であり, 無作為の探索では, 探索効率が低下することが予測される. ここでは次の前提のもとで探索の効果の定式化を試みる.

(1) 前提条件

- データベース内に探索可能なノードが N 箇所存在するものとする.
- そのうちのちょうど 1 箇所に被探索情報が存在する.
- 被探索情報が第 i ノードに存在する確率を P_i とする.

(2) 探索の効果

N 箇所のノードからなるデータベースにおける被探索情報源の存在位置の曖昧さは, エントロピー H を用いて,

$$H = -\sum_{i=1}^N P_i \log P_i \quad (4.1)$$

と表される. ここで, データベース内のノードを n 箇所探索した場合, この探索により, 曖昧さが減少し, H' となる.

$$\Delta H = H - H' \quad (4.2)$$

とすると, エントロピー減少 ΔH は n 箇所の探索の効果と考えられる. これ以降, 本章では ΔH の期待値 $\overline{\Delta H}$ を, 探索によって期待される効果と考える.

4.2.2 EBS 法

$\overline{\Delta H}$ を単純に最大にするには全てのノードを 1 回に探索してしまえばよいが, その選択は効率的探索の現実的な解ではない. 一般に, 情報探索の実際の効率は, その探索コストと探索時間を考慮した上で評価されなければならない. また, 実際の情報探索には, 一般に探索方法が複数あり, それぞれが 1 回の探索効果や探索コストにおいて特徴を持っている. 従って, 情報の探索方針の決定は次の行為の選択になる.

1. 1 回に何箇所のノードを探索するか.
2. どの方法を使うか.

各回の探索方針 (探索方法, 探索ノード数を含む) を p , p による探索コストを $C(p)$, p によるエントロピー減少の期待値を $\overline{\Delta H(p)}$ として, ここでは探索行為の効果に関する評価

量 E を

$$E = \left(\frac{1}{C(p)} + \alpha \right) \overline{\Delta H(p)} \quad (4.3)$$

と考える.

$\overline{\Delta H(p)}$ は探索の期待効果を表す. 従って, 1 回の探索効果が大きいほど探索が終了するまでの時間が短いため, $\overline{\Delta H(p)}$ は探索時間の尺度となる. 一方, $\overline{\Delta H(p)}/C(p)$ は単位コストあたりの探索期待効果を表し, この値が大きい程探索コストの減少が期待できる.

従って, これら 2 つの項の和が探索コストと探索時間の双方を考慮した探索期待効果の評価となる.

評価量 E では, α により, 評価の重みを変えることが可能である. すなわち, α を大きくとり, 評価量最大の探索を繰り返し実行すると探索時間優先探索, 小さくとると同様にして探索コスト優先探索をそれぞれ行うことができる.

この E を最大にするような p を選択する方法を EBS (Entropy Based Search) 法と呼ぶ.

なお, 本来, 上で提案した EBS 法は, データベース内のちょうど 1 箇所ノードに被探索情報が存在する場合に最適性を目指した手法であり, これ以降は, この手法に基づいた情報探索法について議論する. ただし, 実際の探索システムでは被探索情報が複数存在する場合, あるいは全く存在しない場合もありうる. このような場合については第 5 章で述べる.

4.2.3 各ノードにおける被探索情報の存在確率

先に述べた, 各ノードにおける被探索情報の存在確率 P_i は, 各ノードに存在する情報についての以下の 2 つの性質からおおまかに推定する.

1. 情報の類似性
2. 過去の探索経験

まず 1 については, あるノード A において, ある情報 a が存在するとき, a と類似する情報 a' も, ノード A に存在する可能性が高いと推定しうるということである. 各ノードに蓄積される情報が個人ユーザ自身によって編集されたものであるとすると, 類似する情報が集中する傾向が強いと考えられる.

2 については, 現在から t だけ過去のある時点で, ある情報 b をあるノード B で発見した経験がある場合, $1/t$ と相関のある確率で現在も情報 b がノード B に存在すると推定しうるということである. 各ノード内の情報がオーナーであるユーザによって書き換えられる割合は, 時間の経過に従って大きくなるのは明らかである.

本章で対象とする具体的システムにおける P_i の計算方法は 4.5.2 システムの構成で述べる。

4.3 画像情報探索方式

ここでは、4.2章で述べた超分散データベースにおける情報探索法である EBS 法の画像情報探索への適用法を述べる。

4.3.1 画像情報の探索手法

一般に、画像情報探索の手法としては、

- キーワード探索 [20]
- 印象語探索 [21][22]
- ラフスケッチによる探索 [23] [24]
- 特徴量を用いた類似画像探索 [25] [26]
- ブラウジング

などが提案されている。しかし、これらの探索は信頼度およびコストの面で一長一短であり、探索先ノードにおける被探索画像情報の存在確率が推定できる場合には、その探索の状況により使い分けられるのが理想的である。

本章では、低信頼度低コスト探索として、キーワード探索、高信頼度高コスト探索としてブラウジング、双方の中間の探索として特徴量を用いた類似画像探索の3つを用いる。

(1) キーワード探索

画像情報は、言語情報と比較して遥かに多義性・曖昧性が高いため、探索先のノードに探索者の意図した画像が存在しない場合にも存在するという結果を返す可能性があり、探索の信頼度は低い。ただし簡便な探索であるためコストも低い。

(2) ブラウジング

探索者の意図通りの画像を探索でき、探索の信頼度が高い反面、ブラウジング用データを送信するなど、探索のコストも高い。

(3) 特徴量を用いた類似画像探索

画像の特徴を表す属性を用いて、キー画像に類似する構造を持つものを探索する。一般には、色、ヒストグラム、テクスチャ、領域構成、形状情報、およびそれらの組み合わせな

どがある。探索の信頼度，コストともに上の(1)，(2)の中間と考えられる。本章では，画像の隣接するブロック同士の色情報を探索キーとする手法を用いる。この手法については4.4で触れる。

4.3.2 探索手法の仮定

ここでは，本章で用いる探索手法について3つの仮定をおく。

(1) 探索の対象となる画像情報に関する仮定

探索の対象となる画像情報を自然画像に限定する。

(2) 探索に要する時間に関する仮定

探索するノード数に関わらず，同時に1種類（キーワード探索／特徴量探索／ブラウジングのいずれか）の探索を行う際に経過する時間は全て一定であると仮定する。

(3) 探索の精度に関する仮定

先の4.3.1で述べた3種の探索に対して探索先ノードは，被探索情報が存在するという応答 (yes)，あるいは，存在しないという応答 (no) のどちらかを返信するとして，次の仮定をおく。

1. キーワード探索

(a) 探索先ノードに被探索情報が存在するとき

- yes を返信する確率 : 1
- no を返信する確率 : 0

(b) 探索先ノードに被探索情報が存在しないとき

- yes を返信する確率 : \hat{p}_l
- no を返信する確率 : $1 - \hat{p}_l$

2. 画像の特徴量を用いた探索

(a) 探索先ノードに被探索情報が存在するとき

- yes を返信する確率 : 1
- no を返信する確率 : 0

(b) 探索先ノードに被探索情報が存在しないとき

- yes を返信する確率 : \hat{p}_m
- no を返信する確率 : $1 - \hat{p}_m$
- ただし, $0 < \hat{p}_m < \hat{p}_l < 1$

3. ブラウジング

- 探索誤りは無い

ブラウジングを用いれば, 探索先ノードにおける被探索情報の有無が明らかとなるが, その他の探索では被探索情報が実際には無い場合にも yes を返信する可能性があることを仮定しているため, 単独では完全な探索が行えない.

従ってここでは, 高信頼度探索まで含んだ次の4つの Method を考え, この Method の選択に EBS 法を適用し, 評価量 E が最大となる方針で毎回の探索を行うものとする. このアルゴリズムについては次の 4.3.3 および 4.3.4 で述べる.

4.3.3 探索手法の組み合わせ

まず, 3段階の探索手段を組み合わせた, 4通りの探索のオルターナティブを作成する. 各オルターナティブはパラメータ n を持つ.

- Method 1

1 回に探索するノード数を n として, n 全てにブラウジングのみを用いる.

- Method 2

n 全てに最初に画像の特徴量を用いた探索を用い, yes を返信してきたノードのみに, ブラウジングを用いる.

- Method 3

n 全てにキーワード探索を行い, yes を返信してきたノードのみに, ブラウジングを用いる.

- Method 4

n 全てに最初にキーワード探索を用い, その中で yes を返信してきたノードのみにさらに画像の特徴量を用いた探索を行い, これにも yes を返信してきたノードのみにブラウジングを行う.

4.3.4 探索アルゴリズム

本方式の探索アルゴリズムについてまとめる。

1. 各ノードにおける被探索情報の存在確率 (P_i) を推定する。
2. 未探索ノードを情報存在確率 (P_i) の大きい順に並べ、大きい方から n 個 Method 1~4 のいずれかを用いて探索するものとする。各方式について E を計算して、 E を最小とする n と $Method$ を決定し、これに従って探索を行う。
3. 被探索情報が探索したノード中に、
 - (a) 存在した場合 探索終了
 - (b) 存在しなかった場合

P_i を次のように変換して 2 へ戻る。まず探索済みのノードについては、 $P_i = 0$ とする。次にそれ以外のノードの P_i については次式の P'_i とする。

$$P'_i = \frac{P_i}{\sum_{k=1}^N P_k} \quad (4.4)$$

以上のような方式を用いて、本章では実際の画像探索を行うシステムを設計・製作した。このシステムの詳細については 4.5 で述べる。

4.3.5 解析

実際に適当な離散確率分布を仮定して、その確率分布に従うデータベースの探索効率を解析した。

(1) 仮定した離散確率分布および定数値

第 i ノードに情報が存在する確率 P_i を、

$$P_i = \frac{1}{i \cdot \sum_{k=1}^N \frac{1}{k}} \quad (4.5)$$

とした。すなわち $i = 1$ のノードにおいて存在確率は最大で、ノード番号 i の増加に伴い、単調減少する。

探索コストの値は、

- 低信頼度探索 C_i : 10

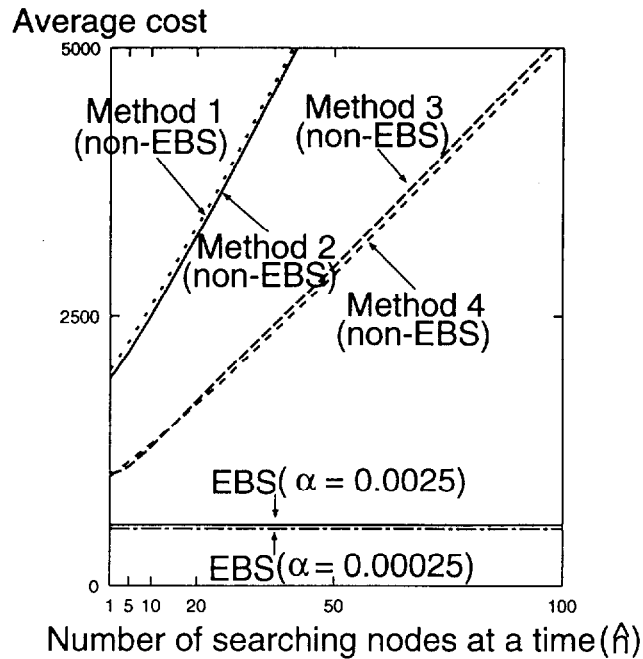


図 4.1: 平均探索コスト

- 中信頼度探索 C_m : 80
- 高信頼度探索 C_h : 100

また、探索先ノードに被探索情報が存在しない時に yes を返信する確率は、

- 低信頼度探索 \hat{p}_l : 0.4
- 中信頼度探索 \hat{p}_m : 0.2
- 高信頼度探索 : 0

さらに総ノード数 N は 100 とした。

(2) 解析結果と考察

ここでは、EBS 法を用いずに Method 1 から 4 について単純に、1 回の探索ノード数を一定にした場合と、本章での提案方式とを比較した。

図 4.1 および図 4.2 に解析の結果を示す。

このグラフの横軸は、両図とも Method 1~4 について、1 回にまとめて探索するノード数 \hat{n} を示している。縦軸は、図 4.1 が被探索情報を発見するまでにかかる探索コストの期待値で、図 4.2 が同様に探索時間の期待値である。

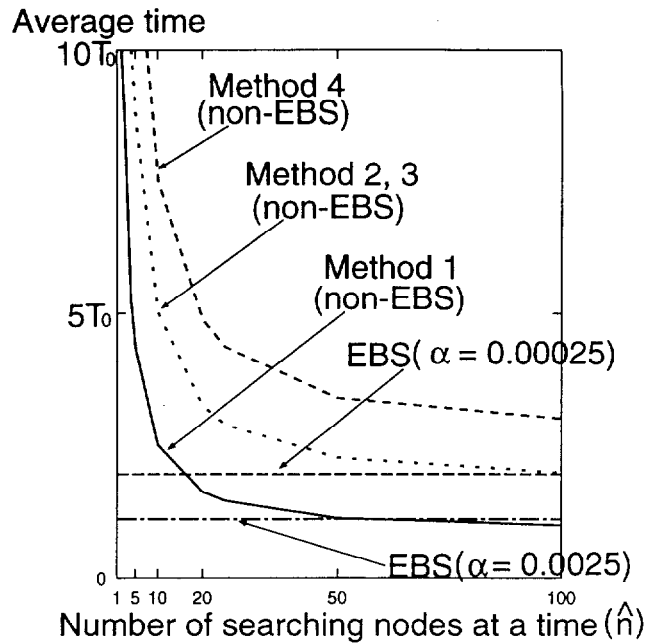


図 4.2: 平均探索時間

\hat{n} が大となるに従い、探索コストの期待値は増加し、探索時間の期待値は減少する。

また、Method 1, 2 と Method 3, 4 でコストに開きが出るが、これは、Method 3, 4 では、コストが比較的小さい低信頼度探索を用いているためであり、この直線の傾きの違いは、 C_h , C_m , C_l の比に依存する。

提案方式では、EBS 法に従い、1 回の探索ごとに Method 1, 2, 3, 4 のいずれの手法で存在確率の高い順に何箇所のノードを探索するのが最も効率が良いか、先に提案した評価関数を用いて判断しているため、横軸の値は決まらない。従って、図 4.1, 4.2 とも、横軸に平行な直線で表した。 α の値については、その探索が、探索コストと探索時間のどちらに重みをおいているのかによって、適当に設定できる。

図 4.1, 4.2 から、エントロピーの減少量を考慮せず、1 回の探索数を一定にしたまま行う探索と比べて本方式では、コスト的にも時間的にもほぼ良好な結果が出ていると言える。

4.4 色対情報を用いた類似画像検索手法

あるノードへ探索キーが与えられた場合、そのノード内で、探索キーとマッチする情報の有無を確認する作業を、本章では検索と呼ぶことにする。ここでは、提案探索方式に適

用するための画像の特徴量を用いた検索手法として、Goh らによって提案された色対を用いた検索手法 [27] を用いる。

4.4.1 色対による検索方式

上の手法では、ある画像を一定のサイズのブロックに分割したあと、その中のあるブロックの色（例えば出現頻度の一番高い色）を隣接するブロックの色と組み合わせて対を作る。この手法では、この「対」の考え方を基に、人間の視覚に最も適合したマンセル表色系を用い、照明の影響を受けにくい色相属性の対を利用した検索法を実現している。実際の検索時間を削減するために、画像データベースの各データに対して前処理を行ない、色対ファイルを作る。検索を行なう時、すでに作成してある色対情報を持つ色対ファイルを用いる。色対ファイルのサイズは原画像よりはるかにデータ量が少ないので、高速な画像検索が可能となる。

4.5 実験システム

提案方式を実際に搭載したシステムを設計・製作した。

4.5.1 システムの動作と探索対象情報

本システムで目的とする動作は次の通りである。

1. ユーザは目的画像を表すキーワードとキー画像 (類似画像) を入力する。
2. システムは情報の類似性および過去の探索経験をもとに EBS 法を用いて複数の候補画像をユーザに見せる。
3. ユーザがこれをもとに判定を行い、目的の画像を発見したところで探索を終了する。

なお、探索対象情報は、具体的には風景・植物・昆虫等の画像情報、数百枚を用いた。

4.5.2 システムの構成

本システムは機能的に、PDB(Personal DataBase) と VDB(Virtual DataBase) の 2 つの部分から構成される。(図 4.3)

- (1)PDB(Personal DataBase)

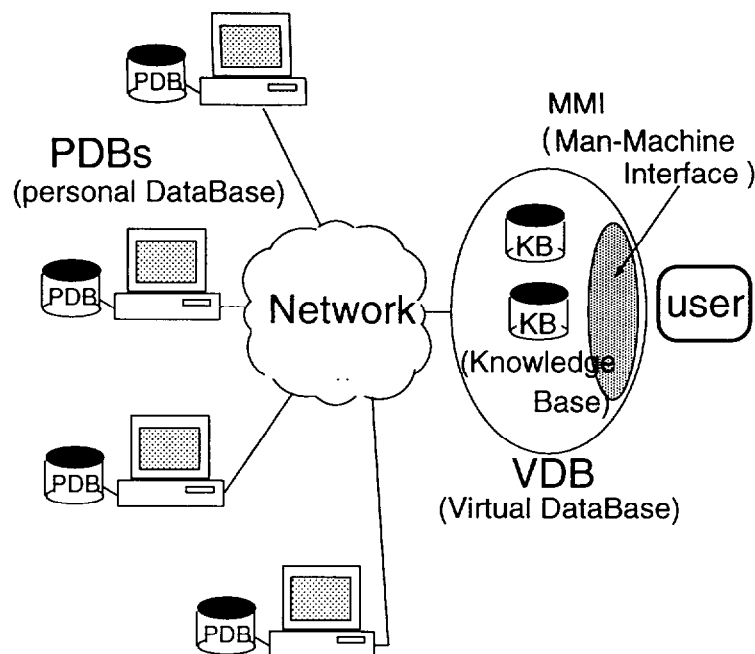


図 4.3: システム概念図

PDB は、ネットワークに接続しているユーザが、自分の所有する画像情報を他のユーザに対して公開した部分である。公開している画像情報は、それぞれ複数のタイトルと色対検索用のデータを持ち、VDB からの探索に応じて yes, no を返信する。さらにブラウジング用のデータも合わせて持ち、VDB からの要求に応じて提供する。

(2)VDB(Virtual DataBase)

VDB は、ユーザに対して情報探索サービスを提供する部分である。実際には MMI(Man Machine Interface) と KB (Knowledge Base) の 2つの部分から構成される。MMI は、ユーザと本システムとの接点であり、ユーザからのキーワード入力およびキー画像入力、ユーザへの縮小したブラウジング画像および発見した被探索画像の提示を行う。また、KB は先に提案した、EBS 法に基づく探索機能を持つ。

EBS 法で用いる初期の被探索情報存在確率 P_i は、次の手法により求める。

(a) 類似度グラフおよび hyperlink の作成

KB は MMI から渡される情報をもとに PDB(被探索ノード) 内の被探索情報の存在確率を推定して探索方式を決定する。本システムではキーワードのみを用いて存在確率を推定することにした。

1. KB 内部に、キーワード (タイトル) 間を意味的類似性により接続したグラフ (類似度

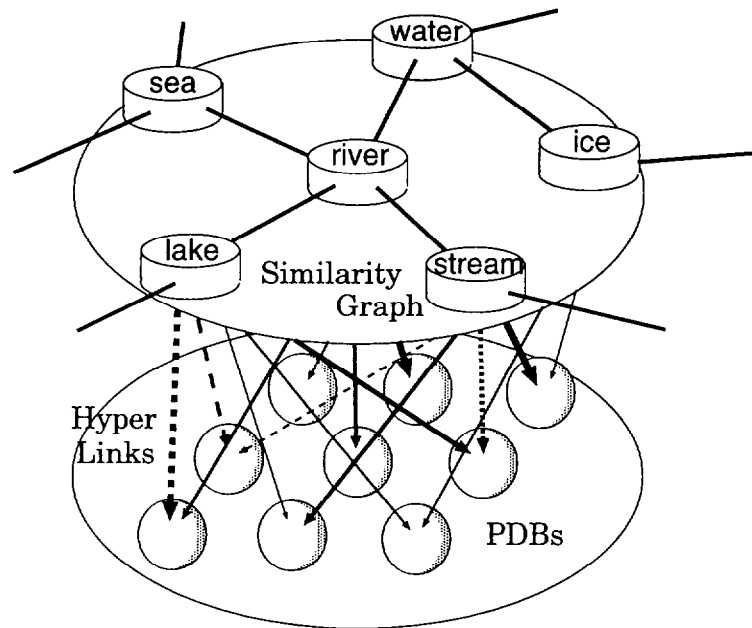


図 4.4: 類似度グラフとハイパーリンク

グラフ)を持たせる。(図 4.4 上部分)ただしこの意味的類似性は分類語彙表 [28] を参照して規定した。

2. 各キーワードは、過去に探索経験のある複数の PDB と hyperlink により KB 内で仮想的に接続される。(図 4.4 下部分)
3. 各 hyperlink は参照される頻度に応じて link の重みを変化させ、使用されない link はエイジングにより削除されるようにする。
4. PDB からの yes, no の情報に基づいて、グラフの hyperlink の重みを変化させるフィードバック機構を持たせる。

(b) 被探索情報存在確率の計算

本システムでは KB 内で常に次のような処理を行う。グラフ上のタイトルである $node_j$ からある PDB_i への hyperlink の重みを $W_k(j, i)$ とする。この $W_k(j, i)$ は、前回の探索により被探索情報の存在が確認された直後で 100 とし、そこから時間の経過により単位時間につき 1 ずつ 0 になるまで減少させ、0 で hyperlink を切断する (次回の探索まで 0 のままとする)。

新たな探索の開始に伴い、本システムではさらに次のような処理を行う。ユーザの投入したキーワードを意味する類似度グラフ上のノードと、 $node_j$ との意味的類似度を $W_g(j)$ とし、その2ノード間のグラフ上の最小ホップ数が、

- 0 および 1 の場合 : $W_g(j) = 100$
- 2 および 3 の場合 : $W_g(j) = 10$
- 4 以上の場合 : $W_g(j) = 0$

のように定める。(ただし、最小ホップ数が0のノードとは、キーワードを意味するノードそのものを指す。)

ここで、 PDB_i における被探索情報の存在確率 P_i を $W_k(j, i)$, $W_g(j)$ をもとに次のように定める。ただし、類似度グラフ上の全ノードに対応するノード番号の部分集合を M , PDB の総数を N とする。

$$P_i = \frac{\sum_{j \in \{M | W_g(j) \neq 0\}} \{W_g(j) + W_k(j, i)\}}{\sum_{l=1}^N \sum_{j \in \{M | W_g(j) \neq 0\}} \{W_g(j) + W_k(j, l)\}} \quad (4.6)$$

(3) 実験システムの構成

製作された画像探索システムは、ワークステーション (Sun Sparc Station) 上のソフトウェアとして実現した。

PDB は、複数のサイト間の10台のワークステーション上に置いた。サイト間の一部はATM回線で接続し、通信回線による遅延を最小限にした。

4.5.3 実験結果

図4.5は、実際に稼働しているMMIの表示画面例である。右上のウィンドウは、画像の特徴量(色対)検索に用いるキー画像で、この画像に近い画像を探索する。右下がブラウザで、キーワードと特徴量(色対)検索の両方の手法により各PDBから探索した被探索画像情報の候補を表示している。ユーザのブラウジングにより最終的に目的の画像が発見された後、その画像情報を該当PDBよりダウンロードしたものがMMI本体の下部分のウィンドウ内の画像である。

以上のようにして製作された画像探索システムは、おおむね良好な性能を示している。しかし、以下のように問題点も残されている。

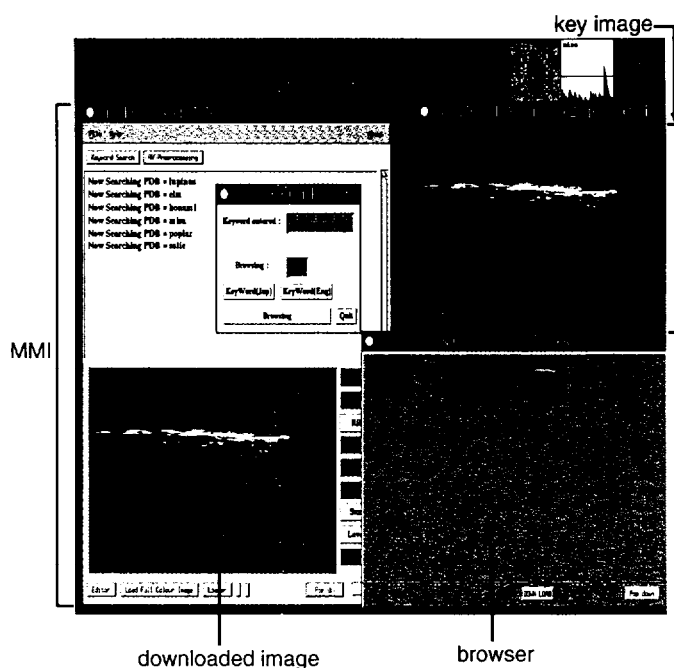


図 4.5: MMI の表示画面例

1. 類似度グラフを作る際に、キーワード間の類似性を正確に求めることは困難である。
2. システムの立ち上げ直後のような、探索経験が全く無い場合の探索は、無作為の探索になる。

これらの問題点については今後、検討する予定である。

4.6 まとめ

本章では、超分散データベース環境における画像情報探索システムの提案とその実験結果について述べている。具体的には、エントロピーに基づいて毎回の探索における探索データベース数と手法の組み合わせを決定する新しい方式および特徴量として色情報を用いた類似画像検索の新しい方式について提案した。さらに、これらを用いた画像探索プロトタイプシステムの設計法と実験結果を述べている。

今後は作成したシステムの規模を拡大してより広範囲の実験を行うと同時に、超分散データベース問題の前提を拡張していく予定である。例えば本章では、被探索情報は全ノードのうちの1箇所が存在するという前提をおいているが、これを拡張して複数箇所に存在す

る場合と、1箇所も存在しない可能性がある場合に対応できるように探索方式を改良していく。

第 5 章

画像情報の効率的探索手法

5.1 まえがき

本章は問題をネットワーク内の画像等の大容量情報探索に限定して、より一般的な前提の上でその最適探索方式を検討したものである。

以下の 5.2 では超分散データベースの定義とその探索問題の定式化を行い、5.3 では探索戦略について述べると共に、提示した探索戦略の最適性を検証する。5.4 では 5.3 で提案した探索方式に関して特性を解析し、探索情報の誤差が与える影響について考察する。5.5 では、計算量を考慮した実際的な探索のための方式の簡略化を提案し、その妥当性について述べる。

5.2 超分散データベースとその探索問題

5.2.1 超分散環境に散在する情報源

ここではインターネットにおける WWW のようにネットワーク内に存在するデータベース内の情報探索を対象とする。情報が画像探索である場合、次のような特徴がある。

[特徴 1]: 各データベース（以後、ノードと呼ぶ）から画像を探索者に送ってきてはじめてそれが目標画像（ターゲット）か否か判断できる。画像情報が大容量であることを考慮すると、1 回の画像送信には大きな通信コストがかかる。

[特徴 2]: 探索者側の探索装置（以後、サーチャと呼ぶ）では一般に並列処理が可能であるが、一度に処理できる量には制限がある。制限はサーチャに接続されている通信路容量、サーチャの同時表示可能枚数、処理能力等に起因する。

以上の特徴より本章では超分散環境における画像処理問題に対して、一般性を失うことなく次のような仮定をおく。

[仮定 1]: 探索コストとして、探索に要する画像通信コスト、探索に要する時間を考える。

[仮定 2]: サーチャは複数のノードに対して同時探索が可能である。各ノードには一般に複数の候補画像があるが、ここでは簡単化のため各ノードの候補画像は1枚とする。サーチャが目標ノードにアクセスすると、各ノードからサーチャに1枚の候補画像を送信し、探索者はその1枚がターゲットか否かを判断するものとする。

[仮定 3]: 探索に要する通信コストはノードとサーチャの距離、ノードから送信される画像の情報量により定まるが、ここでは各画像の情報量は同一であり、コストは距離にのみ依存すると考える。この結果、総通信コストは距離比例定数にノードから送信される画像の枚数を乗じた値となる。

[仮定 4]: 探索時間コストは探索回数に比例するものとする。なおここで1回の探索とは、複数個のノードにサーチャが同時アクセスして、送信されてくる候補画像がターゲットか否かを判断する過程を指す。

[仮定 5]: 各ノードにターゲットが存在する確率は既知であるものとする。

なお実際の探索にあたっては、サーチャが各ノードにキーワード等を送信し、これをもとに各ノードが複数の候補画像を返信するような手続きが多くなる。本章では各ノードの候補画像を1枚と仮定しているが、送信される画像の情報量をノードによって異なるように設定することにより、複数の候補画像がある場合に拡張できる。

5.2.2 探索問題

超分散データベースの探索問題は次のように定義できる。

「以上で述べた特徴および仮定のもとに、平均総時間コストと平均総通信コストの荷重和からなる平均探索コストを最小とするような探索方式を求める。」

本章では、具体的に事前確率に関する知識のもとで、平均探索コストを最小とするような、同時探索ノード数およびその探索該当ノードを決定する方法を求める。

なお、一般には特徴2よりサーチャの探索能力には限りがあり、同時探索可能ノード数には制限があるが、ここでは同時探索可能ノード数は十分大きいとして議論する。

5.3 探索の戦略

探索は複数の探索行為の系列となる。1回の探索時に、サーチャは事前知識およびそれまでの探索経緯に基づいて探索行為（具体的には同時探索ノード）を決定する必要がある。

しかし、その探索により、ターゲットが発見できる場合とできない場合があり、最終的なコストの期待値を求めることは困難である。そこで、本章ではその探索行為のみから最終総コストを推定し、これが最小となるよう探索行為を決定する方式を採用する。なお、この探索方式は各探索時点で最良の方法を選択することを試みるため、最急降下法 (steepest descent) あるいは貪欲戦略 (greedy algorithm) のひとつとなる。

5.3.1 探索方式

不定標的の場合、探索による結果はターゲットを発見できるか否かだけであり、発見できなかったことで、他のノードでの存在確率が増加することはない。従ってある探索行為 p により探索されるノードの集合を $S(p)$ とすると、この探索によりターゲットを発見する確率 $P(p)_{sum}$ は、

$$P(p)_{sum} = 1 - \prod_{i \in S(p)} (1 - P_i) \quad (5.1)$$

となる。ここで、探索行為 p に要する通信コストの和を $C(p)_{sum}$ として、探索終了までに必要な通信コストの予測値を $C(p)_{sum}/P(p)_{sum}$ 、同様に、探索終了までに必要な時間コストの予測値を $T_0/P(p)_{sum}$ とする。これは同一の能力を持つ探索を継続した場合の平均必要コストである。従って、モデル 1 と同様に重み W_1 , W_2 を用いて、総探索コスト予測値 $E(p)$ は、

$$E(p) = (W_1 \cdot C(p)_{sum} + W_2 \cdot T_0) \cdot \frac{1}{P(p)_{sum}} \quad (5.2)$$

と書ける。よって、この値が最小となる探索行為 p を実際の探索に採用することを以って、本論文で提案する探索の方式とする。

5.3.2 最適性の検証

ここでは、 $W_1 = 0$, $W_2 = 1$ (時間コスト重視), $W_1 = 1$, $W_2 = 0$ (通信コスト重視) の 2 つの場合について提案する探索方式の最適性を証明する。

まず時間コストのみを考慮する探索では、一度に探索するノード数に制限が無いとすると、最小探索回数を 1 回として全ノードを同時に探索した場合、最小のコストで探索が終了する。よって、この探索が最適探索である。

一方、 $W_1 = 0$, $W_2 = 1$ の場合、(5.2) 式は以下のようなになる。

$$E(p) = \frac{T_0}{P(p)_{sum}} \quad (5.3)$$

(5.1) 式より、 $P(p)_{sum}$ の最大値は $S(p)$ が全ノードの集合と一致する場合であるため、 $E(p)$ は全ノードを 1 回に同時探索するという p を選択した場合最小となる。

従って、時間コスト重視の場合、探索方式は最適探索と一致する。

次に通信コストのみを考慮する探索では、最適方式について次の定理が成り立つ。

<定理>

第 i ノードを探索した場合の通信コストを c_i 、そのノードにターゲットが存在する確率を P_i とすると、 c_i/P_i の昇順に 1 箇所ずつ探索を行うと、平均コストは最小になる。

<定理終わり>

定理の証明については付録 A で詳述する。

ところで、(5.2) 式において、 $W_1 = 1$ 、 $W_2 = 0$ の場合は、

$$E(p) = \frac{C(p)_{sum}}{P(p)_{sum}} \quad (5.4)$$

となる。この $E(p)$ を最小とする探索方式が、以上で述べた平均通信コスト最小の方式と一致することを示す。

まず、第 j ノード n_j 、第 k ノード n_k において、

$$\frac{c_j}{P_j} \geq \frac{c_k}{P_k} \quad (5.5)$$

なる関係があるとき、この両者の同時探索を $n_j + n_k$ と表すことにすると、 $E(n_j + n_k) > E(n_k)$ となる。(証明は付録 B に示す.)

次に、全ノード n_1, n_2, \dots, n_N について、

$$\frac{c_1}{P_1} \geq \frac{c_2}{P_2} \geq \dots \geq \frac{c_N}{P_N} \quad (5.6)$$

であるとする。

このとき、全ノードのうち n_1 から n_n までの $n (\geq 2)$ 個について、 $E(n_1 + n_2 + \dots + n_n) > E(n_n)$ であることが導かれる (証明は付録 C に示す)。

ゆえに全ノード中の部分集合である、いくつかのノードを組み合わせると同時に探索するよりも、それらのノードの中で c_i/P_i 最小のノード 1 箇所を探索する方が (5.4) 式のコストの予測値は小さい。よって、全ノード中で c_i/P_i 最小の 1 ノード n_{min} を考えたとき、あらゆる組み合わせのノードを同時に探索する場合のコストの予測値の中で、 $E(n_{min})$ が最小である。

以上のことより通信コストの平均が最小となる最適探索と、提案方式において $W_1 = 1$ 、 $W_2 = 0$ とした探索は、結果的に一致する。

よって、本探索方式は $W_1 = 0$ 、 $W_2 = 1$ 、あるいは $W_1 = 1$ 、 $W_2 = 0$ の場合に最適性を有することが明らかになった。

5.4 特性解析

本章では、不定標的問題であるモデル2について、前章までに提案した探索方式の特性および各ノードにおけるターゲットの存在確率の誤差が探索結果に及ぼす影響の感度について考察する。

5.4.1 提案方式の特性解析

前章で述べたモデル2の総探索コスト予測値((5.2)式)を最小とする探索を繰り返すことが、有効であることを数値解析を用いて検証する。

まず、総ノード数 N は 20 とした。また、各ノードにおけるターゲットの存在確率は、以下の図 5.1 と (5.7) 式の離散確率分布で与えた。ただし、第 i ノードでのターゲットの存在確率を P_i とする。

$$P_i = \frac{1}{i \cdot \sum_{k=1}^N \frac{1}{k}} \quad (5.7)$$

(5.2) 式において、時間コストは、 T_0 : 100 とした。また、通信コストは、80~120 (平均 100) の一様分布として、この組み合わせを 100 セット作り、各々について、(5.2) 式を最小とする探索を繰り返した際のコストの平均を求めた。

また、比較のため、本方式とは別の探索方式として、あらかじめ 1 回に探索するノード数を固定して (1, 2, 4, 5, 10, 20 ノード / 1 回)、ターゲットの存在確率の高い順に探索する方式で同時に計算機上で解析を行った。

図 5.2~5.4 が、通信コストのみの比較、時間コストのみの比較、および両者を加えた総コストの比較である。

探索の効果について比較してみると、今回与えた離散確率分布の場合、総コストの比較では、探索数固定の方式を用いると、1 回に 4 ノードずつ探索を行うのが効率が良いことが解る。しかし、提案した方式を用いた場合には、さらに良い結果を導き、特に重み比 W_2/W_1 の値を 3.5 前後にして用いると、最も良い結果を導くことが解る。

よって提案した探索方式を用いた場合、時間コストあるいは通信コストのどちらかを考慮せずに行うような特別な場合の探索だけでなく、両者のバランスをとったような探索においても十分効果的であることが示された。

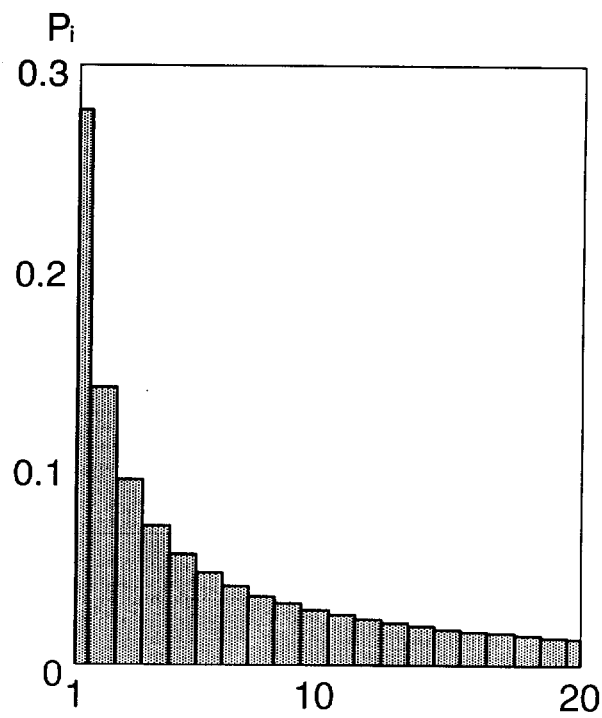


図 5.1: 離散確率分布

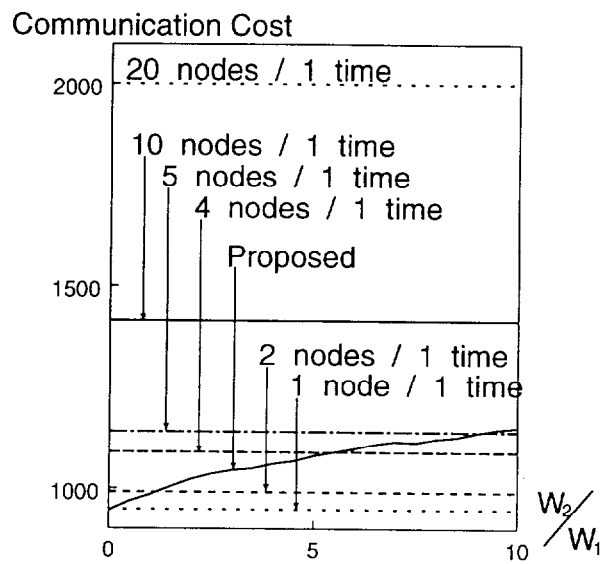


図 5.2: 重み比と通信コスト

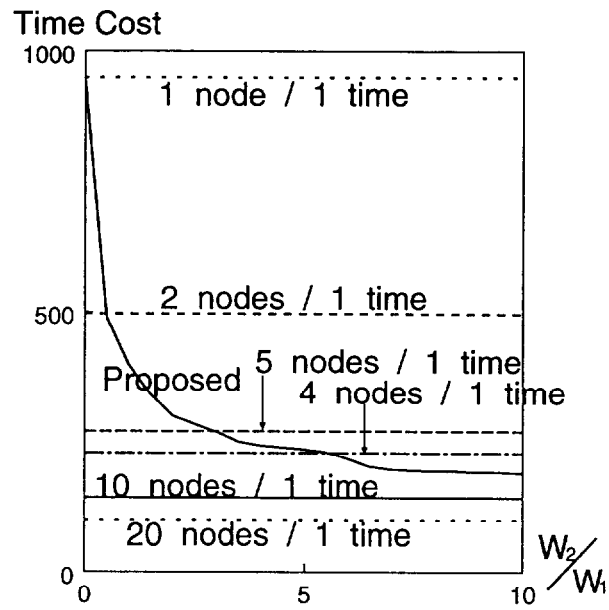


図 5.3: 重み比と時間コスト

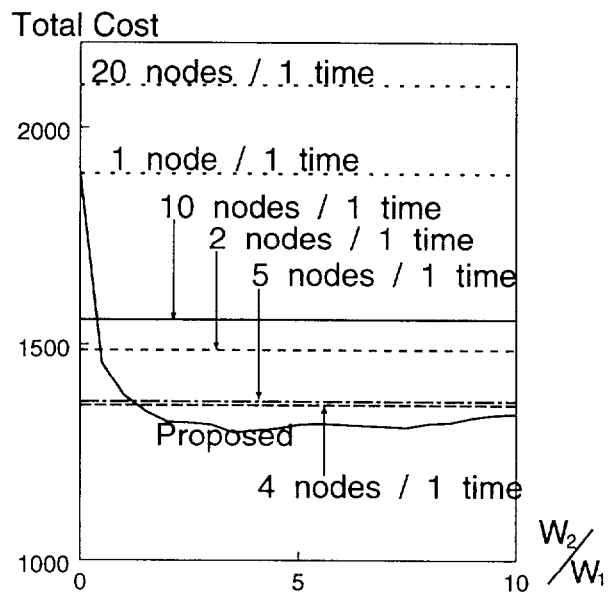


図 5.4: 重み比と総コスト

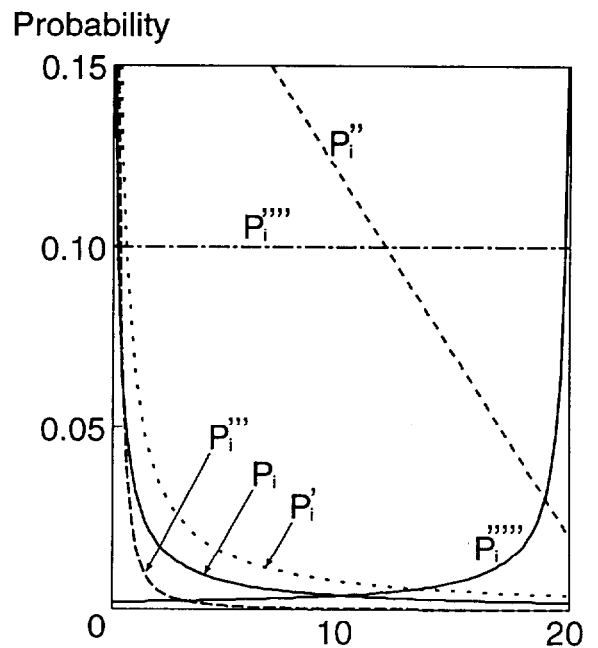


図 5.5: 誤差を含む離散確率分布

5.4.2 ターゲット存在確率の誤差感度

本章で提案する探索方式は、各ノードでのターゲットの存在確率 P_i が既知であることを前提としている。

存在確率は、様々な条件から大まかに推定されるものであるため、一般には誤差を含むことは避けられない。従って、ここでは提案方式の誤差の影響を検討する。

そこで、計算機による解析で、その検証を行った。

(1) 数値解析の方法

仮に探索対象となる各ノードでのターゲットの真の存在確率が、離散確率分布 P_i として与えられるとする。これに対して、実際は誤った離散確率分布をもとに探索を行った場合に、正しいものを用いた場合と探索期待コストにおいてどれだけの差が生じるかを検証する。

本章では、真の存在確率として、図 5.1 と式 (5.7) で示した離散確率分布 P_i を用いた。また、誤った離散確率分布として以下の 5 つを用いた (図 5.5 にグラフを示す)。

$$P'_i = \frac{2}{i \cdot \sum_{k=1}^N \frac{1}{k}} \quad (5.8)$$

$$P_i'' = -0.01 \cdot i + 0.22 \quad (5.9)$$

$$P_i''' = \frac{1}{i^2 \cdot \sum_{k=1}^N \frac{1}{k^2}} \quad (5.10)$$

$$P_i'''' = 0.1 \quad (5.11)$$

$$P_i''''' = \frac{1}{(N+1-i) \cdot \sum_{k=1}^N \frac{1}{k}} \quad (5.12)$$

正しい離散確率分布と、5種の誤った離散確率分布の計6種について、それぞれ通信コストを100通りずつ用意して、本研究で提案した探索法に基づき、期待コストの平均値を計算した。なお、総ノード数は20、通信コストは80から120までの値をとり、平均100の一様分布とした。また、評価関数 $E(p)$ における重み W_1 , W_2 の値も変えて、通信コスト、時間コストの双方にかかる重みの影響も考慮した。

(2) 数値解析結果

図5.6は探索に要したコストの平均値について、正しい離散確率分布を知識として用いた場合と、誤った離散確率分布を用いた場合の比較である。

正しい離散確率分布 P_i と、それに近い傾向の分布である P_i' , P_i'' および離散確率が単調減少する P_i''' はほぼ同様のコストを得ることになるが、全く存在確率がわからない場合である P_i'''' や、実際とは全く逆の情報を与える P_i''''' のような確率分布を用いた場合はコストが大きくなる。

よって本章で提案した探索方式では、各ノードにおけるターゲットの存在確率に関する知識に多少の誤差があっても、真の存在確率とほぼ同傾向の確率が得られれば、それほど大きな影響は受けないと言える。

5.5 方式の簡略化

これまでに述べた、(5.2) 式を最小とする提案方式（以後、方式Aと呼ぶ）は、 $W_1 = 0$ あるいは $W_2 = 0$ の場合は簡単となるが、一般的には計算量が定数のノード数乗のオーダーだけ必要となり、ノード数が多くなると適用困難になる。ここでは以下に示すような方式の簡略化を考え、この方式（以後、方式Bと呼ぶ）が方式Aと比べ、特性低下が小さいことを示す。

方式Bにおいて、被探索ノードの組み合わせは以下の手順で決定する。

1. 探索の組み合わせは、未探索ノード数が M の場合、同時探索数が1の場合から M の場合までの M 通りとする。

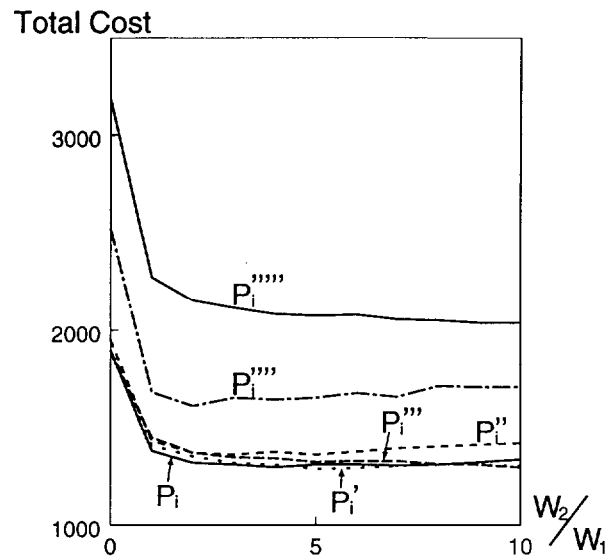


図 5.6: 確率分布に誤差を含む場合の総コスト

2. 各ノードでの通信コストと存在確率の比 c_i/P_i を求め、昇順に並べる。
3. 選択される優先順位をこの順に定め、仮に k 個のノードを選択する組み合わせでは、この優先順位に従って k 箇所のノードを決定する。
4. 以上のようにして、決定された M 通りの組み合わせの中で、総探索コスト予測値 ((5.2) 式) の値が最小のものを毎回の探索で選択する。

本章の探索方式では、方式 A, B ともに評価関数 $E(p)$ を用いて、探索候補を決定している。方式 A では未探索のノード集合の全ての部分集合（ただし空集合は除く）を同時探索されるノード組合せの候補と考え、それぞれの候補について $E(p)$ の値を比較し、最小の値をとるものを探索する。これに対して、方式 B では、未探索ノード総数を M とすると、全ての部分集合ではなく、そのうちの M 通りの組合せを候補として $E(p)$ の値を比較し、最小の値をとるものを探索する。 M 通りの決定方法は、同時探索されるノード数が 1 の場合、2 の場合、3 の場合、 \dots 、 M の場合の M 通りとし、それぞれの場合のノードは、全ての未探索ノードの中から c_i/P_i の昇順に選択される。すなわち、方式 A が全ての未探索ノードの組合せについて $E(p)$ を比較しているのに対し、方式 B はそのうちの一部だけについて $E(p)$ を比較して、探索ノードを決定している。

方式 B を用いる場合、計算量は並べ替えに要する分だけを考えればよいことになるため、多項式時間内での計算が可能である。

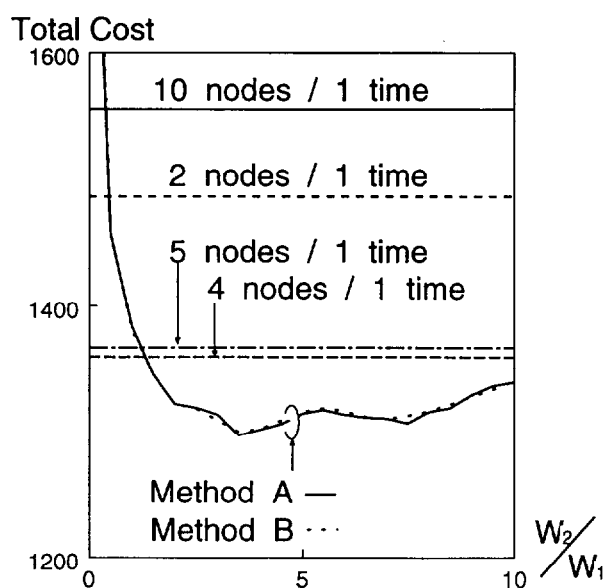


図 5.7: 方式 A および B を用いた総コストの比較

方式 A(Method A) を用いた場合と、方式 B(Method B) を用いた場合、および 5.4.1 の比較用の探索方式を用いた場合の平均コストを比較した結果を図 5.7 に示す。

図 5.7 を見て明らかのように、本章で先に提案した探索方式に忠実にしたがった場合 (方式 A: Method A) の平均コストと、上で述べた探索方式の簡略化を行なった場合 (方式 B: Method B) のそれにはほとんど差が見られないことがわかる。よって実際の探索においては、方式 B を用いることが十分可能である。

5.6 まとめ

本章では、超分散環境に散在する画像情報の効率的な探索の方式について検討を行った。特に目標画像の存在する数が予め解らない場合を対象として、新しい評価尺度をもとにした探索方式を提案して次のことを明らかにした。

1. 提案方式は、通信コストあるいは探索時間のみを評価尺度にした場合、最適性を有する
2. 提案方式は、各ノードでの目標画像の存在確率が推定可能の場合、探索通信コストおよび時間コストを小さくできる

3. 提案方式は目標画像の存在確率推定値に含まれる誤差に対しても強い
4. 計算量を考慮し、方式の簡略化を行った場合でも、探索の特性低下は小さいため、実際にはこちらの方式を用いることが十分可能である

本章では、ネットワーク内での探索ルートは検討対象としていなかったが、今後は本提案方式を適用した場合の探索ルートの最適化を検討するとともに、具体的なシステムに実装していく予定である。

第 6 章

エージェントを用いた探索

6.1 まえがき

本章では、ネットワーク上の情報を対象としたエージェントによる情報探索の効率化を目的とする。まず、探索に要する通信・処理コストと時間コストを定義して、情報探索問題を定式化する。次に、エージェントを数（単一／複数）、機動性より分類して、それぞれの場合についての効率的な情報探索方式を論じる。

以下 6.2 では本研究で扱うエージェントについて述べ、6.3 では情報探索問題を定義し、6.4 では通信ネットワークに関するエージェントモデルの分類を行う。6.5 では、基本モデルに対する最適解を求め、6.6、6.7 ではこれを拡張して各モデルに対する探索方式を求めている。

6.2 エージェントの概要

6.2.1 エージェント型探索の特長

現在 Web 上においては、既にテキスト情報を基本とした検索システムが多数実用化されている。これらのサーチエンジンと呼ばれるシステムは何らかの方法でキーワードと URL からなるインデックス情報を事前に収集しこれをもとにユーザの求める情報を発見するものである。この手法は、ネットワーク全体を巡回し、手元の蓄積情報を更新し続けることによって、ネットワーク内に存在する情報の内容と存在位置を常に正しく把握しようとするものである。いわば、超分散データベースであるネットワークを情報内容と存在位置が整理された通常のデータベースとして扱えるようにしたものと言える。

本章では、超分散データベースの探索に、利用者の要求に応じて探索を行うエージェントを用いる方法を対象とする。本章で対象とするエージェントはサーチエンジンと異なり、事前に情報収集を行わないため、不特定の Web サイトを全て探索する場合には適さず、

- (1) 特定グループのデータベースの中から必要情報を探索する
- (2) 複数のサーチエンジンの中から必要情報を探索する
- (3) サーチエンジン等で与えられた複数の候補サーバから必要情報を探索する

ことに適していると考えられる。

エージェントを用いた探索の特長は以下にあげられる。

(i) ユーザが探索を依頼するだけで、それ以降の探索処理はエージェントが全て独立して行えるため、ユーザ側が用いるシステムとは独立したユニットとして扱える。

(ii) 複数のエージェントを用いる場合、探索対象となるノードを、そのエージェントの近隣に限定できるため、探索のためのデータ送受信によるネットワークへの負荷を抑制することが可能である。特に、探索対象のデータが画像データのような大容量である場合に効果的である。

(iii) 機動性を持つエージェントである場合、自分自身がノード間を移動した方が、画像データのような大きなデータを処理のために送信する必要がなくなり、ネットワークに対する負荷を軽減できる。

本章ではこれ以降、サーチエンジンのような情報事前蓄積型探索とは別のアプローチである、エージェントを用いた探索手法について検討する。

6.2.2 エージェントの定義

本章で扱うエージェントは、情報探索機能を持つ。また、自律性を持ち、ユーザの意図する情報（ターゲット）の探索要求を受けると、これを単独あるいは複数のエージェントが協調して探索を自動的に実行する。最終的には、ユーザに対して探索結果（ターゲット情報または探索失敗の通知）を送信して探索を終了する。

ここで、ターゲットの発見を確認する方法として以下の2つのモデルが考えられる。

- (1) ターゲットをエージェント自身が完全に判別することができる場合。
- (2) ターゲットの確定をユーザが毎回行う必要がある場合。

さらに(2)の場合は、2つのケースが考えられる。

(i) エージェントとユーザの間の距離が近い（機動性が無く、エージェント数が1の場合）、ユーザに確認をするための送信データサイズが画像のサムネイルのように小さく、確認のためのコストが無視できるケース。

(ii) それが可能でないケース

本章では、(1) の場合、および (2) 場合の (i) のケースについて考察する。

6.3 情報探索問題

本章では、これ以降、多数のノードとリンクからなるネットワークを仮定し、このネットワーク上の複数のノードに、情報源が存在すると考える。特に、探索の対象となる情報源が存在するノードを探索ノード、探索者がエージェントを置いて探索の拠点とする（機動性を持つエージェントの場合は移動を開始する）ノードをベースノードと呼ぶことにする。

エージェントは探索開始時点でベースノード上にあり、そこから探索ノードに対して、探索情報（ターゲット）の有無についての確認を開始する。その場合のコストとして通信・処理コストと時間コストを考える。また、ノード間の回線容量、各ノードでのターゲットの存在確率について、以下のような仮定をおく。

6.3.1 通信・処理コスト

探索に要する通信コスト、処理コストを合計して通信・処理コストと呼ぶ。通信コストは次章で分類するエージェントモデルによりその定義が異なる。

(1) エージェントがベースノードから移動しない場合の通信・処理コスト

データの送受信に伴うネットワークへの負荷、およびエージェントが各ノードに存在する情報についてターゲットであるか否かを判断する処理のために生じる計算機への負荷で、通信・処理されるデータサイズと相関を持つコストと考える。

(2) 探索ノードにエージェントが移動する場合の通信・処理コスト

エージェント自身が次のノードに移動する際に生じるネットワークへの負荷、および移動した先のノードに存在する情報がターゲットであるか否かを判断する処理のために生じる計算機への負荷で、エージェント自身のサイズ、および処理されるデータサイズと相関を持つコストと考える。

6.3.2 時間コスト

あるノードにターゲットが存在するか否かを判断するために要する時間を時間コストと呼ぶ。通信・処理コストと同様に、エージェントモデルにより、その定義が異なる。

(1) エージェントがベースノードから移動しない場合の時間コスト

ベースノード，探索ノード間の通信に要する時間，および探索ノードに存在する情報がターゲットであるか否かを判断するために要する時間からなる。

(2) 探索ノードにエージェントが移動する場合の時間コスト

エージェントが探索を終了したノードから次のノードに移動する時間および，移動した先のノードに存在する情報がターゲットであるか否かを判断するために要する時間からなる。

6.3.3 ノード間の回線容量

インターネットのような大規模ネットワークにおいて，ノード間を接続する際の回線容量は，常に変動し，また，ルーティングの変更などもしばしば起きるため，厳密には瞬間ごとの回線容量に変動が生じる。しかし，この時間変動を正確に捉えることは困難である。

本章では，探索時における回線容量等が仮定でき，かつ探索時間内での変動が小さい場合を前提とする。

6.3.4 各ノードにターゲットの存在する確率

各ノードにおいてターゲットの存在する確率を正確に推定することは容易ではないが，現在のサーチエンジンでも独自の方法でヒット率等を求め，提示している。筆者らも既に過去の探索経験および情報の類似性から存在確率を推定する方法を提案している。また，画像情報探索の場合には事前に簡単な特徴量探索によりターゲット画像と候補画像の類似度を求める方法も数多く提案されている。以上より本章では，

- (1) エージェントが過去の探索経験より自ら存在確率を求めうる
- (2) サーチエンジン等から候補サーバを与えられるときに，類似度から存在確率を推定しうる

のいずれかを前提として，ターゲットの存在確率は既知であると仮定する。なお，筆者らの同種の探索方式における研究においては，探索効率は存在確率の絶対値にあまり依存せず，存在確率の大きさの順序の影響が大きいことが明らかになっており，本章で扱う問題においても存在確率の数値に対する感度はそれほど大きくないと予測している。

以上の仮定を用いて，本章では，探索問題を

「探索に要する平均総時間コスト，平均総通信・処理コストの荷重和を最小とするようなエージェントによる探索方式を求める」

と定義する。

6.4 エージェントモデルとその分類

現在，エージェントについては多くのモデルが提案されている。本章では，ネットワーク上の情報探索問題を考える上で，エージェントモデルを以下の2点に着目し，分類を行った（表 6.1）。

- エージェントの数 … 問題解決を単体で行うか，複数で協調して行うか。
- 機動性 … エージェントを実現するソフトウェアの実体が，探索先のノードに移動する能力を持つか否か。

ただし，ここでは，機動性を有するエージェントは探索を行う際にエージェント自らが直接ノードに移動してターゲットの探索を行うものとし，それ以外の探索手段は持たないものとする。

表 6.1: エージェントの分類

	機動性なし (Fixed)	機動性あり (Mobile)
単体 (Single)	(1)	(2)
複数 (Multi)	(3)	(4)

- (1) ベースノードに固定された1エージェントが各探索ノードにアクセスする。(Single-Fixed)
- (2) 1個のエージェントが，複数の探索ノードを巡回する。(Single-Mobile)
- (3) ネットワーク上に複数のベースノードとそれに固定されたエージェントが存存し，各々のエージェントが探索ノードにアクセスする。(Multi-Fixed)
- (4) 1箇所のベースノードから複数のエージェントが出発し，各々が複数の探索ノードを分担して巡回する。(Multi-Mobile)

本章では、以上のモデルのうち (1)~(3) について効率的な探索アルゴリズムの提案とその有効性の検証を行う。(4) のモデルについては、(2) と (3) の組み合わせとなるため、本章では省略する。

また、(1) から (3) のモデルについての実際の応用分野としては、以下のようなものが考えられる。

- (1) 全てのエージェント型探索の基本に相当する。小規模なシステムにより前処理なしに情報探索を行う場合には有効である。
- (2) セキュリティの問題もあり、インターネット全体で利用できるようにすることは困難であるが、あるユーザグループが多数のサイトに分散して配置され、頻繁に更新されるような情報を共有することにしたような場合には有効である。
- (3) マルチメディアデータの探索エンジンにおいて、事前に情報を蓄積するために、複数のエージェントが協調するシステムとして、応用が可能である。既に平岡らによってプロトタイプのプロトタイプ提案、設計、構築がなされている [29]。

6.5 Single-Fixed Type モデル (基本モデル)

6.5.1 探索問題

機動性を持たず、かつ単体で目的を達成するエージェントの場合、エージェントは1箇所のベースノードから、ターゲットを発見するまで、存在する可能性のある探索ノードに順次アクセスすることになる。

この場合、ネットワーク上においてベースノードと探索ノードのみを考えると、ベースノードを中心とした仮想的なスター状のネットワークトポロジにおける情報探索と考えることができる (図 6.1)。

単体のエージェントの場合、

- N_d : 探索ノードの総数
- d_i : i 番目に探索する探索ノード
- p_i : d_i に情報が存在する確率
- c_i : d_i を探索する時に要する通信・処理コスト
- T_i : d_i を探索する場合の時間コスト

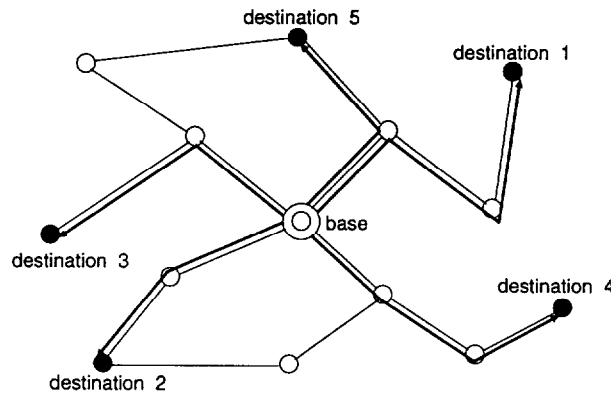


図 6.1: 単体・機動性なしのエージェントの探索

とすると、平均総時間と平均総通信・処理コストの荷重和である評価量 E は次のように求められる。

まず、平均通信・処理コスト E_c は、

$$\begin{aligned}
 E_c &= c_1 + c_2(1 - p_1) + c_3(1 - p_1)(1 - p_2) + \cdots \\
 &\quad + c_{N_d}(1 - p_1)(1 - p_2) \cdots (1 - p_{N_d-1}) \\
 &= c_1 + \sum_{l=2}^{N_d} \left\{ c_l \cdot \prod_{k=1}^{l-1} (1 - p_k) \right\}
 \end{aligned} \tag{6.1}$$

と書ける。次に、平均時間コスト E_t は、

$$\begin{aligned}
 E_t &= T_1 + T_2(1 - p_1) + T_3(1 - p_1)(1 - p_2) + \cdots \\
 &\quad + T_{N_d}(1 - p_1)(1 - p_2) \cdots (1 - p_{N_d-1}) \\
 &= T_1 + \sum_{l=2}^{N_d} \left\{ T_l \cdot \prod_{k=1}^{l-1} (1 - p_k) \right\}
 \end{aligned} \tag{6.2}$$

と書ける。

ここで、探索に要する平均総コストの評価量 E は、 E_c と E_t が同程度の値をとるように正規化した上で、 W_1, W_2 ($W_1, W_2 \geq 0, W_1 + W_2 = 1$) を重みとして用いると、

$$E = W_1 \cdot E_c + W_2 \cdot E_t \tag{6.3}$$

となる。

重み W_1, W_2 は、 E_c と E_t の値を正規化してあるため、探索者の嗜好に応じて自由に決定することができる。これによって通信・処理コストを許容し、時間コストを削減する探索

方針をとるか、その逆の方針をとるか、あるいはその比率をどのようにするかを調整することが可能である。

その際には、ユーザの使用する各計算機・ネットワーク環境における通信・処理コストと時間コストとの均衡点（互いに等価な通信・処理コストと時間コストの比）を何らかの方法で求めることが必要であるが、本章ではこの問題については扱わない。

6.5.2 最適探索法

全ての探索ノード d_j に関し、 $p_j/(W_1 \cdot c_j + W_2 \cdot T_j)$ を計算し、この値の降順に探索ノードにアクセスする方式が最適であることが証明できる。

<証明>

d_1, d_2, \dots, d_{N_d} の順に探索を行った場合の、期待通信・処理コストと時間コストの荷重和の期待値を ε_1 、同様の探索で、特に m 番めの探索と $m+1$ 番めの探索の順序のみを入れ換えた場合の荷重和の期待値を ε_2 とすると、

$$\begin{aligned}
 \varepsilon_1 = & W_1 \cdot c_1 + W_2 \cdot T_1 \\
 & + (W_1 \cdot c_2 + W_2 \cdot T_2)(1 - p_1) \\
 & + (W_1 \cdot c_3 + W_2 \cdot T_3)(1 - p_1)(1 - p_2) \\
 & + \dots \\
 & + (W_1 \cdot c_m + W_2 \cdot T_m)(1 - p_1) \\
 & \quad \times (1 - p_2) \cdots (1 - p_{m-2})(1 - p_{m-1}) \\
 & + (W_1 \cdot c_{m+1} + W_2 \cdot T_{m+1})(1 - p_1) \\
 & \quad \times (1 - p_2) \cdots (1 - p_{m-1})(1 - p_m) \\
 & + \dots \\
 & + (W_1 \cdot c_{N_d} + W_2 \cdot T_{N_d})(1 - p_1) \cdots \\
 & \quad \times (1 - p_2)(1 - p_{N_d-1})
 \end{aligned} \tag{6.4}$$

$$\begin{aligned}
 \varepsilon_2 = & W_1 \cdot c_1 + W_2 \cdot T_1 \\
 & + (W_1 \cdot c_2 + W_2 \cdot T_2)(1 - p_1) \\
 & + (W_1 \cdot c_3 + W_2 \cdot T_3)(1 - p_1)(1 - p_2) \\
 & + \dots \\
 & + (W_1 \cdot c_{m+1} + W_2 \cdot T_{m+1})(1 - p_1)
 \end{aligned}$$

$$\begin{aligned}
& \times (1 - p_2) \cdots (1 - p_{m-2})(1 - p_{m-1}) \\
& + (W_1 \cdot c_m + W_2 \cdot T_m)(1 - p_1) \\
& \quad \times (1 - p_2) \cdots (1 - p_{m-2}) \\
& \quad \quad \times (1 - p_{m-1})(1 - p_{m+1}) \\
& + \cdots \\
& + (W_1 \cdot c_{N_d} + W_2 \cdot T_{N_d})(1 - p_1)(1 - p_2) \cdots \\
& \quad \times (1 - p_{N_d-1})
\end{aligned} \tag{6.5}$$

両者の差を求めると,

$$\varepsilon_2 - \varepsilon_1$$

$$\begin{aligned}
& = (W_1 \cdot c_m + W_2 \cdot T_m)(1 - p_1)(1 - p_2) \cdots \\
& \quad \times (1 - p_{m-2})(1 - p_{m-1})(-p_{m+1}) \\
& + (W_1 \cdot c_{m+1} + W_2 \cdot T_{m+1})(1 - p_1)(1 - p_2) \cdots \\
& \quad \times (1 - p_{m-2})(1 - p_{m-1})p_m
\end{aligned} \tag{6.6}$$

ところで,

$$\frac{p_m}{W_1 \cdot c_m + W_2 \cdot T_m} > \frac{p_{m+1}}{W_1 \cdot c_{m+1} + W_2 \cdot T_{m+1}} \tag{6.7}$$

とすると, $\varepsilon_2 - \varepsilon_1 > 0$ すなわち, $\varepsilon_2 > \varepsilon_1$ となる.

従って, 連続する 2 回の探索では, 探索ノードを $p_j/(W_1 \cdot c_j + W_2 \cdot T_j)$ の降順に探索する場合, 評価量の平均値は小さくなる. このことから探索を終了するまでの評価量の平均値を最小にするためには, 探索順を $p_j/(W_1 \cdot c_j + W_2 \cdot T_j)$ の降順とすればよいことは明らかである.

<証明終わり>

$p_j/(W_1 \cdot c_j + W_2 \cdot T_j)$ を最小とする探索は, 重み $W_1 = 0$, $W_2 = 1$ とすることで時間コスト最小の探索となり, この場合 p_j/T_j の降順の探索となる. また, $W_1 = 1$, $W_2 = 0$ とすると通信・処理コスト最小の探索で, p_j/c_j の降順の探索となる.

以降 6.6, 6.7 では, 上で求めた最適方式を拡張した探索方式を提案する.

6.6 Single-Mobile Type モデル

6.6.1 探索問題

探索を依頼されたエージェントは、ターゲットが存在する可能性のあるノードに順次移動して探索を行っていく。このとき、Single-Fixed Type のエージェントと異なり、ある探索ノードでターゲットが発見できなかった場合には、直接次の探索先に向かうことになる(図 6.2)。また、エージェントはターゲットを発見した時点で、その情報をベースノードに送信して消滅するものとし、それに要する通信・処理コスト、時間コストともに無視できるものとして扱う。

以上から、6.5で述べた評価量 E は次のように求められる。

D : ベースノードと各探索ノードのノード番号の
集合 (ただしベースノードの番号は 0 とする)

$\tilde{c}_{i,j}$: d_i, d_j 間をエージェントが移動し、
探索を行う際に要する通信・処理コスト

$\tilde{T}_{i,j}$: d_i, d_j 間をエージェントが移動し、
探索を行う際に要する時間コスト

ただし, $i, j \in D$

とおくと, d_1, d_2, \dots, d_n の順で巡回する場合の平均通信・処理コストおよび時間コストは,

$$E_c = \tilde{c}_{0,1} + \sum_{k=1}^{N_d-1} \{ \tilde{c}_{k,k+1} \cdot \prod_{l=1}^k (1 - p_l) \} \quad (6.8)$$

$$E_t = \tilde{T}_{0,1} + \sum_{k=1}^{N_d-1} \{ \tilde{T}_{k,k+1} \cdot \prod_{l=1}^k (1 - p_l) \} \quad (6.9)$$

となり,

$$E = W_1 \cdot E_c + W_2 \cdot E_t. \quad (6.10)$$

従って、ここで扱う問題は、(6.10) 式の評価量 E を最小にするような、複数の探索ノードを巡る経路を求めるものとなる。

この問題の最適解を求めるのは困難であり、例えば、 $W_1 = 1.0, W_2 = 0$ として評価量を通信・処理コストのみに限定しても、巡回セールスマン問題を拡張したようなものとなる。巡回セールスマン問題が、すべての探索ノードを通過してベースノードに戻ってくる最短経路を見つけるものであるのに対し、さらにこの問題では、それぞれの探索ノードにターゲット

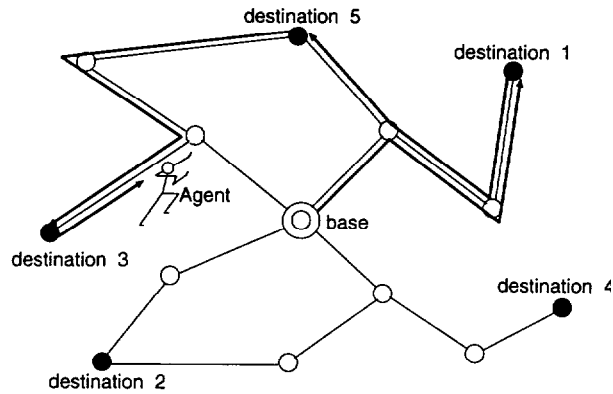


図 6.2: 単体・機動性ありのエージェントの探索

トが存在する確率が与えられており、ターゲットが見つければその時点で探索が終了するという条件がつく。巡回セールスマン問題は NP 完全な問題であることが知られており、明らかにこの問題はより複雑な手順を必要とするため多項式オーダーの最適なアルゴリズムを見つけるのは難しい。

巡回セールスマン問題に関しては、単純貪欲法、大局的貪欲法、挿入法、最小木法、局所探索法など、いくつかの近似アルゴリズムが提案されている [30]。

本章では、先に定義した問題に関して、巡回セールスマン問題の近似法のひとつである挿入法に近い手法を適用し、効率的な探索を試みた。これを提案アルゴリズムとして以下に示す。

6.6.2 探索アルゴリズム

1. 全ての探索ノードおよびベースノードから任意の 2 個 i, j を選ぶ全ての組合せについて、これらのノード間に要する通信・処理コストと時間コストの荷重和の最小値、

$$e_{i,j} = \min[W_1 \cdot \tilde{c}_{i,j} + W_2 \cdot \tilde{T}_{i,j}] \quad (i, j \in D) \quad (6.11)$$

を dijkstra 法などを用いて全て求める。

2. 初期状態として、ベースノードと $p_k/e_{0,k}$ が最大となる探索ノード k を接続リンクで結合する。
3. $p_l/e_{0,l}$ の降順に、ベースノードと接続されていない探索ノード l をリンクで接続されている 2 ノード間に割り込ませ、評価量 E を計算する。リンクが複数ある場合は、全

てについて試みて、最も E を小さくするリンク間にその探索ノードに割り込ませることを決定する。

4. 全ての探索ノードがリンクでベースノードに接続されたところで探索順が決定される。
5. この探索順に従って、ターゲットが発見されるか、全ての探索ノードを巡回し終えるまで、探索を継続する。

提案アルゴリズムは、次の特徴がある。

- 探索ノードの数が $i-1$ 個の場合の探索経路が与えられたとき、探索ノードの数が i 個の最適な探索経路が得られる。しかし、総ノード数が $i+1$ 個以上について最適性が保証されず、最終的に局所最適解で終了する可能性がある。
- 逐次的に求める探索経路は、最終的な最適探索経路に近づくと仮定して近似を行っている。よって、全探索の枝刈りに近い処理を行っている。
- アルゴリズムの計算量 [30] について、探索ノード数を $N(D)$ とすると、
 - (1) の部分では、ベースノードと探索ノードの各々から全てのノードへの最短経路（最小コスト）を求めるので、 $O(N_d^2 \cdot N(D))$
 - (2)~(4) の部分では、まず $p_l/e_{0,l}$ の降順にソートを行うために最悪で $O(N(D)^2)$ 、その後この順に輪の中にノードを割り込ませるために、 $O(N(D)^2)$

となる。よって、明らかに多項式時間で計算は終了する。

上に示した探索経路決定に要する計算処理は、探索ノードが既知であれば事前に行うことができるため、時間コストには含めない、また、ターゲットを特定する処理と比較して計算量も小さいと考えられるため、処理コストにも含めない。

6.6.3 提案アルゴリズムの評価

上で提案した探索アルゴリズムが有効であることを、シミュレーションを用いて検証する。まず、縦 10、横 10 のノード総数 100 の格子状のネットワークを仮定し、探索ノードはそれらのノードからランダムに決定する。今回その個数は 10, 30, 50 の 3 通りとした。

時間コストは一定値とする。また、通信・処理コストについては、処理コストが他のコストと比較して無視できる程度十分に小さいものとし、通信コストは距離と相関があるものとする。

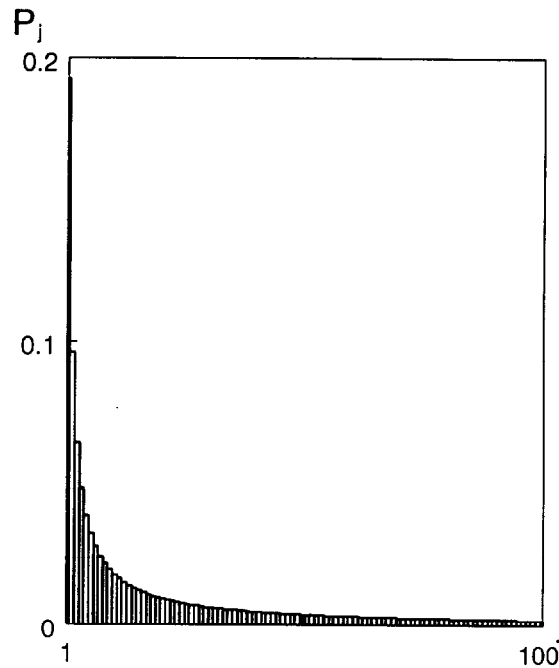


図 6.3: 離散確率分布

各ノード間の接続リンクの通信コストは 80 ~ 120 (平均 100) の一様分布とした。また、探索 1 回当たりの時間コストは全て一定値として、通信・処理コストとのバランスを考慮 (正規化) し、 $\tilde{T}_{i,j} = 400$ と置いた。

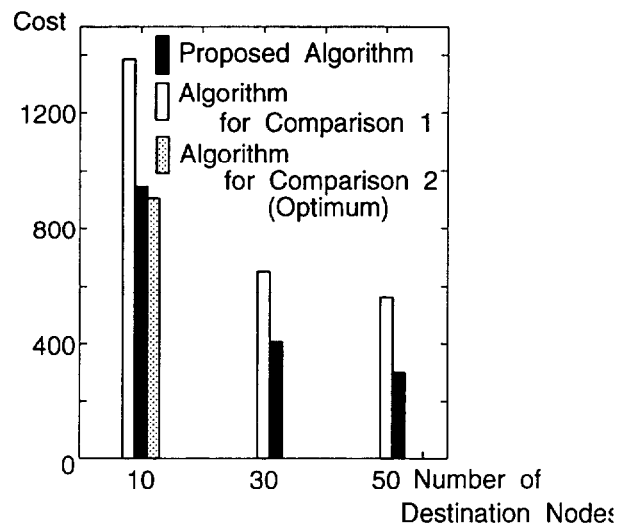
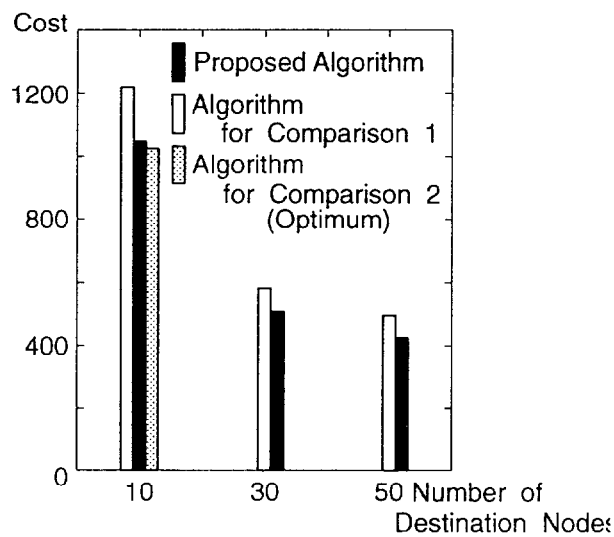
各探索ノードにおけるターゲットの存在確率 p_j は、以下の離散確率分布に基づくものとし、各ノードにランダムに割り当てた (図 6.3)。

$$p_j = \frac{1}{j \cdot \sum_{k=1}^{100} \frac{1}{k}} \quad (6.12)$$

ここでは、どのエージェントが何度めに探索を行っても、各ノードにおけるターゲットの存在確率は割り当てた状態のまま、探索終了まで不変であるものとする。また、エージェントの初期位置はネットワークの中央に固定した。

重みが $W_1 = 1.0$, $W_2 = 0$ と $W_1 = 0.5$, $W_2 = 0.5$ の 2 種類の場合について、以上のような条件で、モデルを生成し、次の 4 種類のアルゴリズムを用いて評価量 E を求める。この処理を 1000 回繰り返し行い、それぞれの場合の平均総コストを求めた。

同時に、提案アルゴリズムと比較するために、以下の 2 つの比較アルゴリズムを用いた。

図 6.4: Single-Mobile Type 総探索コスト ($W_1 = 1.0$, $W_2 = 0$)図 6.5: Single-Mobile Type 総探索コスト ($W_1 = 0.5$, $W_2 = 0.5$)

- 確率順探索 (比較アルゴリズム 1)
- 厳密解を求めるアルゴリズム
(比較アルゴリズム 2)

なお、比較アルゴリズム 2 は、探索する順番の全組合せについて評価量 E を計算し、最小となるものを解とする (最適解を得るアルゴリズム、探索ノードの数が 10 の場合のみ)。

図 6.4, 6.5 に 1000 回の探索コストの平均値を示す。

提案アルゴリズムを用いた場合、ノード数にかかわらず、比較アルゴリズム 1 を用いた場合よりも期待コストが小さくなっている。探索ノードの数が 10 の場合に関しては、提案アルゴリズムは比較アルゴリズム 2 (厳密解を求めるアルゴリズム) と比較してほとんど変わらない特性を示した。

また、どのアルゴリズムも探索するノード数が増えるほど期待コストが小さくなる。これは、ターゲットの存在確率を図 6.3 のような分布として、ランダムに割り当てているため、探索ノードの数が増えるとそれに伴って高い確率をもつノードの数が増すことが原因であると考えられる。

6.7 Multi-Fixed Type モデル

6.7.1 探索問題

機動性がない複数のエージェントを用いて情報探索を行う場合は、6.4 で述べたような固定エージェントがネットワーク上に複数存在し、それら各々が、同一のターゲットを求めてそれぞれの担当する探索ノードにアクセスすることになる (図 6.6)。

具体的には、ここで扱うエージェントの探索メカニズムは次のようなものとする。複数のエージェントは、それらのうちいずれかが情報を発見するか、または担当する探索ノードをすべて探索し終えるまで連続して並行探索を行う。エージェントのいずれかがターゲットを発見した時点で、他のエージェントに対し、その旨を伝達し、全てのエージェントは探索作業を終了する。

ここで、エージェント間通信に要するコストは、情報探索自体に要するコストに比較して無視できるほど十分に小さいものとする。

以上のように前提をおき、扱う問題を、「エージェント数、およびベースノード位置が与えられた場合の、通信・処理コストと時間コストとの荷重和を最小にする探索法 (探索ノード

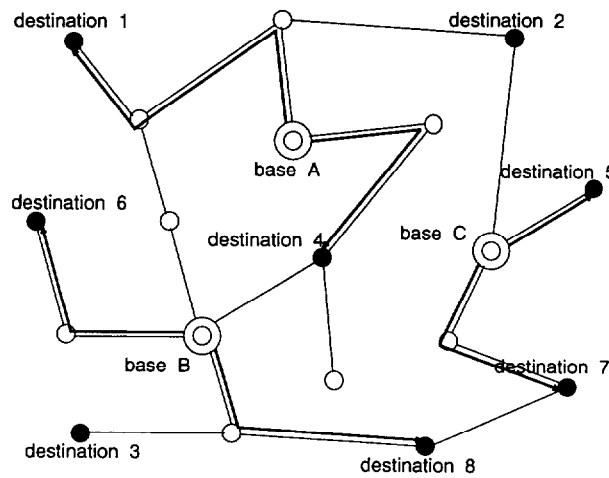


図 6.6: 複数・機動性なしのエージェントの探索

ドの割り当て, 探索順) を求める」こととする. ここでは, エージェント数, ベースノード位置に関しては既知であるとし, エージェントの配置問題に関しては扱わない.

6.7.2 提案アルゴリズム

ここでは, Single-Fixed Type で用いた手法を拡張した以下の2つのアルゴリズムを提案する.

- 基本アルゴリズム (Basic Algorithm)
- 修正アルゴリズム (Modified Algorithm)

各ベースノードに対して, 単なる確率順の探索を行うよりも効率的なアルゴリズムとして, まず, 基本アルゴリズムを考える. この方式は, あるエージェント x から探索ノード y までの探索に要する通信・処理コストおよび時間コストを $c_{x,y}$, $T_{x,y}$ と表すことにすると, $p_j / (W_1 \cdot c_{i,j} + W_2 \cdot T_{i,j})$ の降順に同数の探索ノードを振り分けるものである. さらに基本アルゴリズムを修正した方式として, 修正アルゴリズムを提案する. このアルゴリズムは具体的には, 担当振り分け, 初期スケジューリング, 動的修正の3つの過程より構成される.

(i) 担当振り分け:

$W_1 \geq W_2$ ならば, 各探索ノードは最も通信・処理コスト (距離) が小さいエージェントの担当とする.

$W_1 < W_2$ ならば、探索ノードを確率の降順に、各エージェントに割り振る。ただし、割り振りにあたっては、各エージェントに割り振られたノードの探索時間ができるだけ平均化するようにした上で、可能な限り各探索ノードが距離の近いエージェントに割り振られるように配慮する。

この結果、通信・処理コストを重視した場合には、各探索ノードまでの通信・処理コストを平均化して総探索時間の偏りを許容している。また、時間コストを重視した場合には、各探索ノードまでの通信・処理コストに差が生ずることを許容し、各エージェントに課せられる探索時間の和を平均化している。

(ii) 初期スケジューリング：

それぞれのエージェント i は、基本的には振り分けられた各探索ノード j を $p_j / (W_1 \cdot c_{i,j} + W_2 \cdot T_{i,j})$ の順に1箇所ずつ探索を行う。この処理は、担当する探索ノードが全て探索されるか、いずれかのエージェントがターゲットを発見するまで継続する。

(iii) 動的修正：

ただし次のような場合、例外的に他のエージェントが担当する探索ノードを探索する。

(a) それぞれのエージェント i は、毎回、次に探索する探索ノードを決める際、他のエージェントが担当するものも含め、すべての未探索の探索ノード j について $p_j / (W_1 \cdot c_{i,j} + W_2 \cdot T_{i,j})$ を計算し、最大となる探索ノードを選択する。

(b) 選択された探索ノードが初期スケジューリング上、他のエージェントが担当するものである場合、総コストの期待値を計算し、初期スケジューリングのそれが、修正を行った場合と比較して大きいときに限り、スケジュールの変更を行う。(各ノードでの存在確率が既知であるため、期待値の計算は可能であるが、計算法の詳細は省略する。) これにより、探索候補の衝突が起きた場合には、通信・処理コストの小さい方のエージェントが担当する。

(c) 探索候補に決定した探索ノードを担当できな

かったエージェントは次に $p_j/(W_1 \cdot c_{i,j} + W_2 \cdot T_{i,j})$ が大きい値をとる探索ノードを候補とする。

(iv) 担当する探索ノードがなくなったエージェントは、その時点で探索を終了する。

(v) ターゲットを発見するか、全てのエージェントが探索を終了した場合、この探索は終了する。

6.7.3 アルゴリズムの計算量

基本アルゴリズムと修正アルゴリズムに要する計算量について考察する。ただし、探索ノード数を $N(D)$ 、エージェント数は $N(D)$ より十分小さい定数とする。

(1) 基本アルゴリズムの計算量は、

- 与えられた $T_{x,y}$, $c_{x,y}$ を用いて、各エージェントごとに全ての探索ノードについて、評価量 $p_j/(W_1 \cdot c_{i,j} + W_2 \cdot T_{i,j})$ を計算するための計算量が、 $O(N(D))$
- さらに評価量の降順にソートするための計算量が各エージェントごとに、 $O(N(D)^2)$
- 各エージェントについて、そこから見た評価量の降順に、探索ノードを1つずつ分配して振り分けるための計算量が全部で $O(N(D))$

である。全体としては $O(N(D)^2)$ の計算量に収まるため、ノード数が多数である場合でも、計算コストの面では十分現実的なアルゴリズムであると言える。

(2) 修正アルゴリズムの計算量は、

- (i) においては、 $W_1 \geq W_2$, $W_1 < W_2$ の双方について、全探索ノードを通信・処理コストまたは確率でソートしたのち、各エージェントに振り分けるための計算量が、全体で $O(N(D)^2)$
- (ii) においては、各エージェントの担当ごとに振り分けられた探索ノードを評価量の降順にソートするための計算量が $O(N(D)^2)$
- (iii) においては、動的修正のために、各エージェントから見た未探索の全ノードの評価量を求めるための計算量が $O(N(D))$ 、さらに担当ノード衝突のための例外処理まで含めた探索を、ターゲットが発見されるまで繰り返すために必要な計算量が $O(N(D))$

である。こちらも全体としては $O(N(D)^2)$ の計算量に収まるため、やはりノード数が多数である場合でも、計算コストの面では十分現実的なアルゴリズムであると言える。

6.7.4 探索アルゴリズムの評価

6.6.3と同様に、縦 10、横 10 のノード総数 100 の格子状のネットワークを仮定して、シミュレーションを行った。ただし、探索 1 回当たりの時間コストは一定値とし、通信・処理コストとのバランスをとる（正規化を図る）ため、全て $T_{i,j} = 1600$ とした。

また、エージェントの数は 2、4、6 の 3 通りでモデル網の上に図 6.7 のようにそれぞれ配置した。

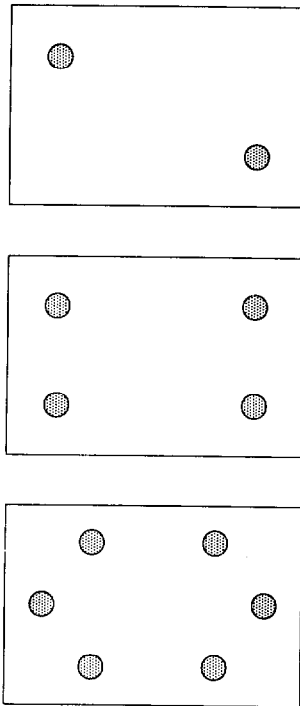


図 6.7: エージェントの配置

以上のような条件で、重みを $W_1 = 1, W_2 = 0$ と $W_1 = 0.5, W_2 = 0.5$ の 2 種類の場合について、基本アルゴリズムと修正アルゴリズムの 2 種類を用いて平均総探索コストを求めるとい処理を 1000 回繰り返し行い、それぞれコストの平均を求めた。

図 6.8, 6.9 に 1000 回の総探索コストの平均を示す。

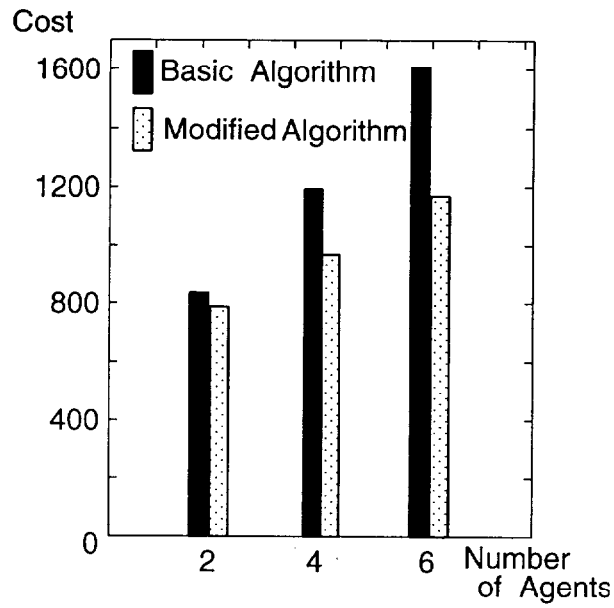


図 6.8: Multi-Fixed Type 総探索コスト ($W_1 = 1.0$, $W_2 = 0$)

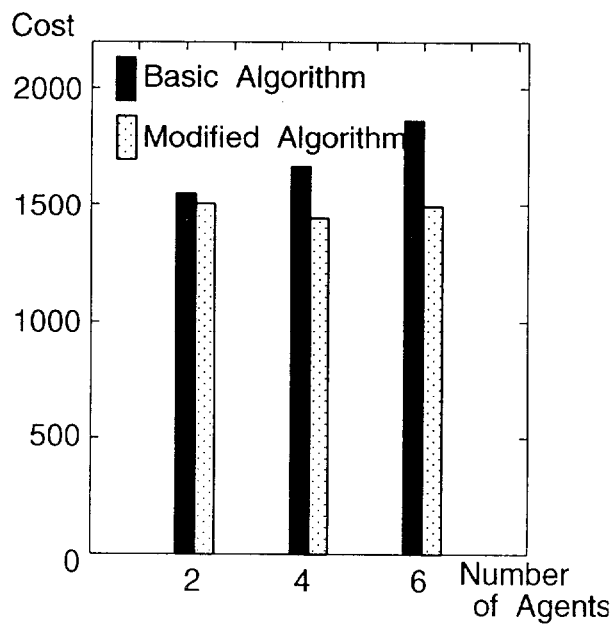


図 6.9: Multi-Fixed Type 総探索コスト ($W_1 = 0.5$, $W_2 = 0.5$)

$W_1 = 1, W_2 = 0$ (図 6.8) と, $W_1 = 0.5, W_2 = 0.5$ (図 6.9) のいずれの場合も 修正アルゴリズムは基本アルゴリズムよりも総探索コストが小さい。

また, エージェント数が多い程, 修正アルゴリズムの改善効果が増している。これはエージェント数が増すと, 1 エージェントあたりの探索ノード数が少なくなり, 基本アルゴリズムのように無理に同数の割り当てを行うことが不適当となることを示している。

エージェント数が増加しても総探索コストが改善していないが, これは, ある程度確かなターゲットの存在確率を得ている場合, 必要以上のエージェントは無駄になることを示しており, ターゲットの存在位置に関する情報量をゼロにした場合はエージェント数が多い方が双方のアルゴリズムにおいて総探索コストが小さくなることを確認している。

6.8 まとめ

本章では, エージェントによる探索問題を, エージェントの数, 機動性という2つの観点から分類した。そして, それらのうちいくつかの場合について, あるノードにターゲットが存在する確率を推定できるという条件のもとで, 効率的な情報探索法を提案し, その有効性を示した。本論文で対象としなかった問題としては,

- エージェントの探索結果をユーザが確認する必要がある, そのためのネットワークコストが無視できない場合
- 複数の機動性のあるエージェントの問題

がある。さらに本研究の解析結果をより完全にするためには, 今後の課題として,

- (1) より高精度なターゲットの存在確率の推定方法の検討
- (2) 評価関数 E における重み W_1, W_2 の合理的な決定手法の検討
- (3) 探索手段として Mobile と Fixed の両方の特徴を持った方式に関する検討
- (4) ノード間のリンクの帯域が時間的に変化する場合を想定した効率的探索手法についての検討
- (5) エージェントが発見した情報がターゲットであるか否かの判断をするために, ユーザに対して確認をするモデルに関する検討

があげられる。

第 7 章

探索打ち切りの効果

7.1 まえがき

これまでの研究では、探索は探索情報（ターゲット）が発見されるか、あるいは探索対象となるネットワーク内の情報源（サーバ）全てにターゲットが存在しないことが確認されるまで継続されるという前提であった。しかし、インターネットのような大規模なネットワークでは、サーバ数が非常に多く、このような前提は現実的ではない。

実際のインターネット上での情報探索を考えた場合でも、繰り返しターゲットを探索した結果発見できない場合には、ユーザは適当なところで探索を打ち切っていると考えられる。限られたコストの中で効果的な情報探索を行う場合、探索の打ち切りを決断する際の、明確な基準を設け、これを考慮した探索を行うことは重要である [31]。

本章では、探索エラーが生じる危険率を一定値以下に抑制する探索回数を計算し、この回数を最大探索回数として探索打ち切りの基準とする手法、および、探索報酬と探索エラーによるペナルティを定義し、探索から得られる利得の期待値を見積もることによって探索者にとって有利な回数まで探索を継続する手法を提案する。

7.2 インターネットにおける情報探索

本章では、インターネット上に散在する WWW サーバを対象として、ユーザにとって必要な情報（ターゲット）の探索を行う場合の、打ち切りを考慮した効率的探索手法について考察する。

このようなインターネット上のサーバの探索について、図 7.1 のようなシステムを前提に以下のような仮定をおく。

- ネットワーク内に散在する探索可能なサーバは N 箇所存在するものとし、これらを

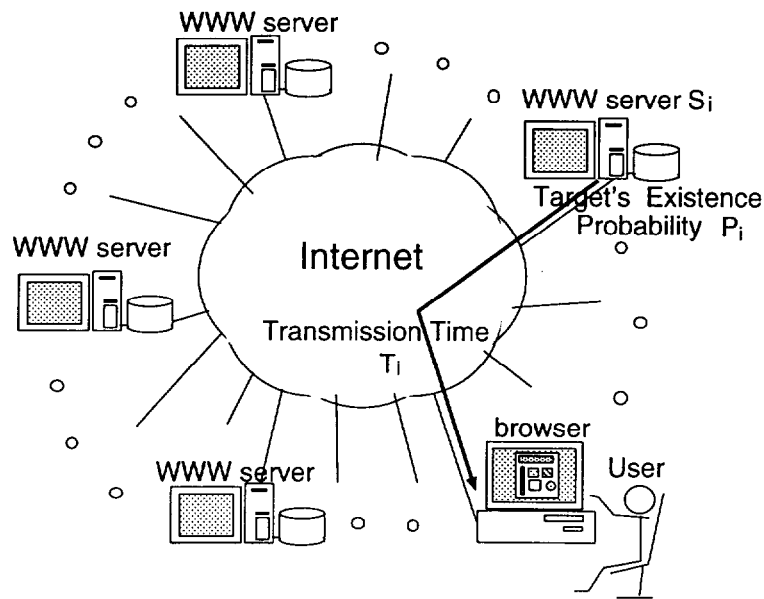


図 7.1: システムイメージ

S_1, S_2, \dots, S_N とおく

- N は十分大きいものとする
- 各サーバ S_i におけるターゲットの存在確率 P_i は既知であり、これらの確率は互いに独立で、他のサーバの探索結果により影響を受けない
- 探索はブラウザを用いて1サーバずつ逐次行い、ターゲットが発見された場合、探索は終了する
- 探索に要するコストは探索開始時点から探索を終了するまでの総時間（各サーバを探索する場合の時間コストの和）のみとし、各サーバ S_i を探索するための時間 T_i は全て既知である

以上のような仮定をもとに、本研究では効果的な探索打ち切りの基準として以下の2つの手法を提案し、考察する。

- 統計的手法を用いた探索打ち切り
- 探索成功の報酬と探索エラーが発生した場合のペナルティを考慮した探索打ち切り

前者については7.3、後者については7.4で述べる。

7.3 統計的手法を用いた探索打ち切り

探索対象となる総サーバ数を N 、最大探索数を r_0 として、この回数までの探索によってターゲットが発見できなかった場合には、探索を打ち切るものとする。また、簡単のため、ここでは各サーバ S_i におけるターゲットの存在確率 P_i は一定値 P_0 と仮定する。この打ち切りによって、 N 個のサーバのうちターゲットを持つものが存在したにもかかわらず、発見できない場合を探索のエラーと定義すると、エラーの発生する確率 P_f は、

$$\begin{aligned} P_f &= (1 - P_0)^{r_0} P_0 + (1 - P_0)^{r_0+1} P_0 + \dots \\ &\quad \dots + (1 - P_0)^{N-1} P_0 \\ &= (1 - P_0)^{r_0} - (1 - P_0)^N \end{aligned} \quad (7.1)$$

と書ける。

P_f を r_0 の関数と考え、 r_0 で微分すると、

$$\frac{dP_f}{dr_0} = (1 - P_0)^{r_0} \log(1 - P_0) \quad (< 0) \quad (7.2)$$

となり、 P_f は r_0 の単調減少関数である。よって、 r_0 を連続値と仮定した場合、ある P_f ($0 \leq P_f \leq 1 - (1 - P_0)^N$) を与える r_0 ($0 \leq r_0 \leq N$) は、一意に決まることになる。

ここで、統計的な危険率 α を用いて、エラーが発生するという仮説を棄却することを考える。ターゲットが存在しながらそれを見逃す確率は α だけ残るが、この危険率を一定の値に抑制することにより、打ち切りを行う探索の精度を保證することができる。

すなわち、

$$(1 - P_0)^{r_0} - (1 - P_0)^N < \alpha \quad (7.3)$$

より、

$$r_0 > \frac{\log\{\alpha + (1 - P_0)^N\}}{\log(1 - P_0)} \quad (7.4)$$

であるから、これを満たす、最小の非負の整数値 r_0 を求めると、適正な探索回数を得られる。

この手法は、本来各サーバによって固有の値をとる P_i を全て一定値 P_0 と仮定している部分において正確さを欠くが、どのサーバにおける P_i も正確に推定することが困難である場合や、平均的な値のみ推定可能であるような場合には有効であると考えられる。

また、次の章では探索によって得られる報酬および探索エラー発生時のペナルティを定義する手法を提案するが、このような定義が難しい場合には有効であると考えられる。

7.4 探索報酬とペナルティを考慮した探索打ち切り

7.4.1 探索報酬とペナルティ

本章では、さらに以下の仮定をおく。

- 探索に要するコストは、ここでは時間コストのみを考え、簡単のため、各サーバを探索するコスト T_i は一定値 T_0 とする
- よって探索は、 P_i の降順に行われる場合、期待コストが最小となるため、この順に行う
- あるターゲットを探索する際に、そのターゲットが得られた場合の報酬を定義し、1回の探索コスト T_0 の定数倍であるとする
- 探索の報酬は、探索者およびターゲットにより固有の値であるとする
- 探索を打ち切ることによりエラーが発生した場合、探索側にペナルティが与えられるものとし、探索の報酬と同様に1回の探索コスト T_0 の定数倍であるとする
- 探索エラーによるペナルティは、探索者およびターゲットにより固有の値であるとする

7.4.2 探索手法

以上の仮定に従って、打ち切りの基準を示し、探索の方針を決定する手法を提案する。

1回の探索に要する探索コストを T_0 、ターゲットを発見することによる報酬を aT_0 、エラーによるペナルティを bT_0 とする。ただし、 a, b は非負の定数とする。

$k (\geq 0)$ 回めの探索までにターゲットが発見できなかった場合、 $k+1$ 回めの探索を行うことにより得られる利得の期待値 E_{k+1} は、

$$E_{k+1} = a \cdot T_0 \cdot P_{k+1} - (k+1)T_0 - b \cdot T_0 \left\{ \prod_{i=1}^{k+1} (1 - P_i) - \prod_{i=1}^N (1 - P_i) \right\} \quad (7.5)$$

と表せる。ここで定数 a, b は探索者およびターゲットにより任意の値をとる。

探索の事前に、 E_1 から E_N までの値を計算することにより、探索を継続してもターゲットが発見されない場合に、最大何回の探索を行うのが探索者にとって有利であるかを知ることが可能である。

本章では、探索者にとって有利な回数で、探索を打ち切る手法を提案する。

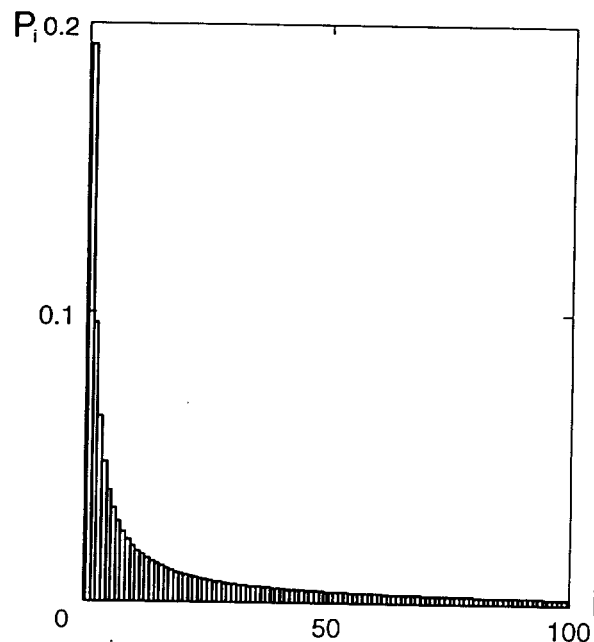


図 7.2: 離散確率分布 その1

7.4.3 数値解析

本手法を用いて探索を打ち切る場合の特性を以下に示す.

各サーバ S_i ($1 \leq i \leq N$) におけるターゲットの存在確率 P_i を,

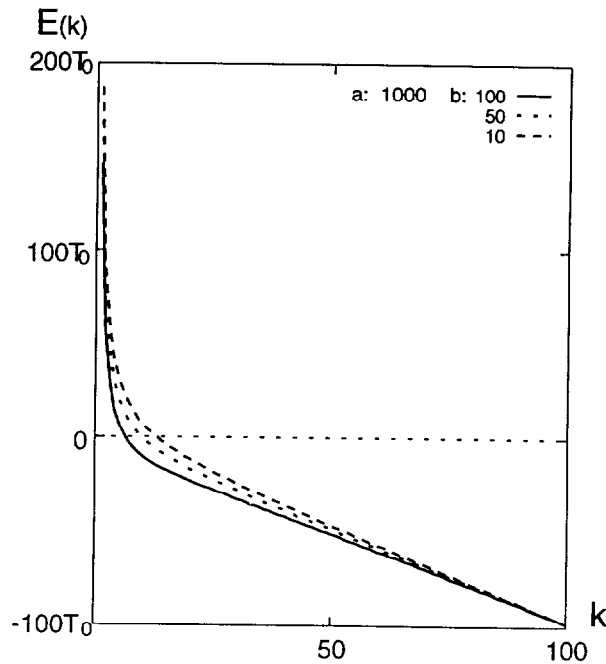
$$P_i = \frac{1}{i \cdot \sum_{j=1}^N \frac{1}{j}} \quad (7.6)$$

のように与える (図 7.2). また, 探索は i の昇順に行うものとする.

総探索サーバ数 N を 100, 定数 a の値を 1000 として, b の値を 100, 50, 10 の 3 通りに変えた場合, k 回で探索を打ち切ったときの探索利得の期待値 $E(k)$ は, 図 7.3 のような結果になる. a の値を 100, 10 とし, b は同様に変えた場合の $E(k)$ を図 7.4 および図 7.5 に示す.

図 7.3 では, ペナルティを規定する定数 b の値が 100, 50, 10 のどの場合でも, k の値が比較的小さい間 (それぞれ $k=7, 9, 13$ まで) は $E(k)$ は正の値をとる. この間は探索をしても利得が期待できることを示している. k の値がこれより大きくなると, $E(k)$ の値は負に転じ, 単調減少する. これは, これ以上探索しても利得を得られる期待が薄いことを示している. よって, この場合は $E(k)$ の値が負に転じる直前までが平均的に利得を得ることが可能な範囲であり, それ以上の探索はすべきではない.

図 7.4 では, 探索の報酬を規定する a の値が図 7.3 と比較して小さいため, ほとんどの場

図 7.3: 探索利得 E と探索回数 k の関係 その 1

合, 探索による利得は得られる期待が薄い. ただし, ペナルティを規定する定数 b が最小の 10 である場合のみ, 最初の 3 回までは利得の期待値は正である.

図 7.5 では, 探索の報酬を規定する a の値がさらに小さくなり, 全ての場合で探索による利得が得られる期待は薄い. $a = 10$ であるため, 特に k が 10 を越える探索にメリットは無い. これらの場合は b の値にかかわらず全て探索を 1 回も行うべきではない.

図 7.3 から図 7.5 の全てについて, k の値を 100 に近付けると, $E(k)$ の値は $-100T_0$ に収束する. これは, 全てのサーバを探索した場合にはエラーは起こらず, かつ, P_i の分布より 100 回めに探索するサーバでのターゲット存在確率が非常に小さいため, 報酬の期待値が 0 に収束するからである.

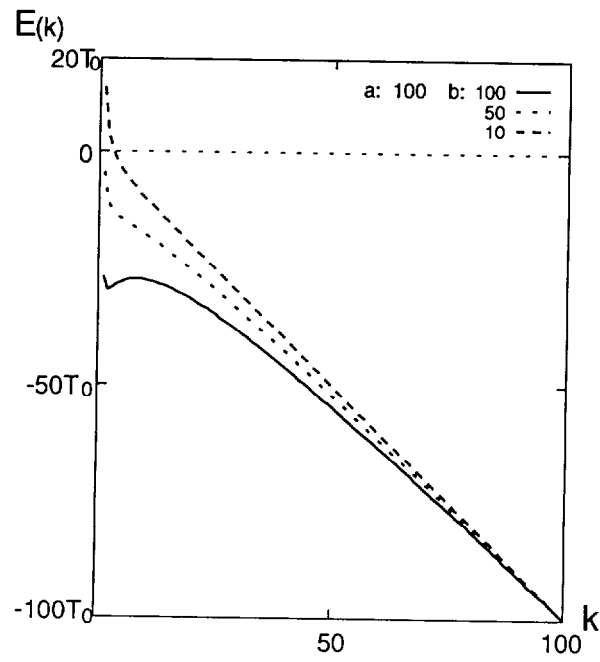
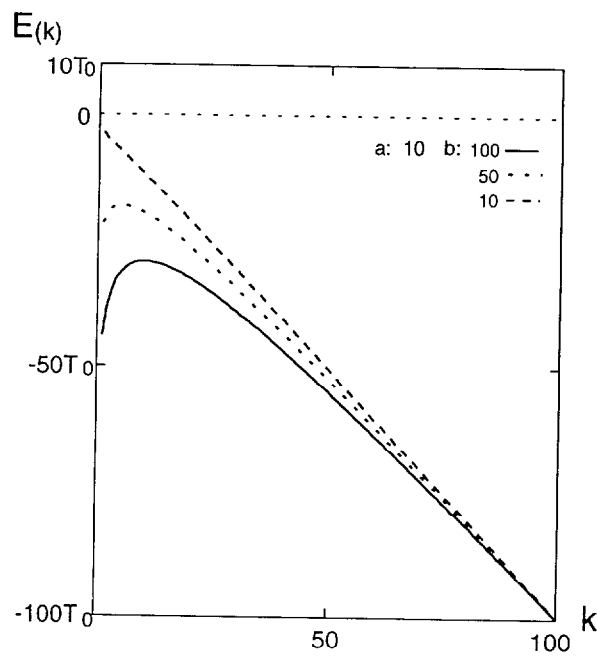
次に各サーバにおけるターゲットの存在確率 P_i を,

$$P_i = 0.2 \quad (7.7)$$

のように与えた場合 (図 7.6) を考える.

先の場合と同様に, 総探索サーバ数 N を 100, 定数 a の値を 1000, b の値を 100, 50, 10 の 3 通りに変えた場合の探索利得の期待値 $E(k)$ を, 図 7.7 に示す. また, a の値を 100, 10 とした場合の $E(k)$ を図 7.8 および図 7.9 に示す.

図 7.7 では, ペナルティを規定する定数 b の値が 100, 50, 10 のどの場合でも, k の値に

図 7.4: 探索利得 E と探索回数 k の関係 その 2図 7.5: 探索利得 E と探索回数 k の関係 その 3

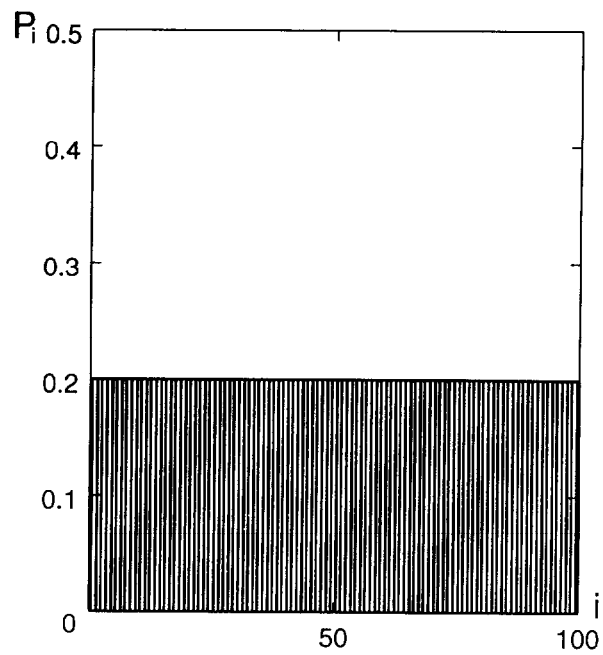
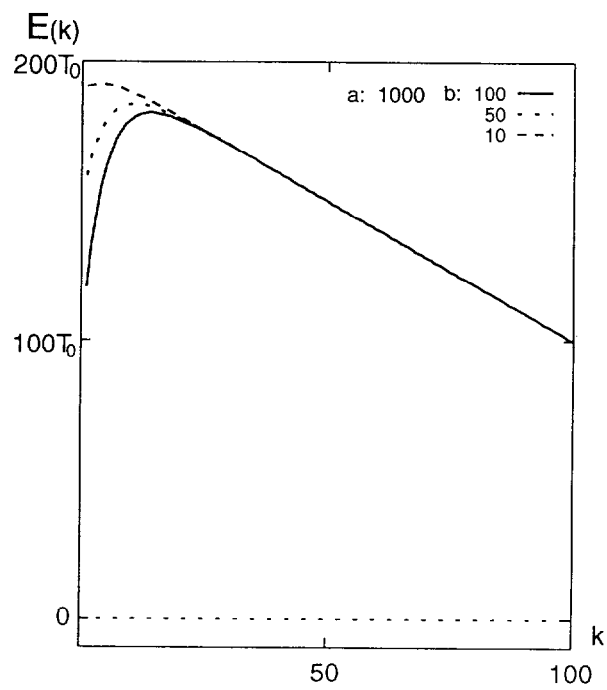


図 7.6: 離散確率分布 その 2

図 7.7: 探索利得 E と探索回数 k の関係 その 4

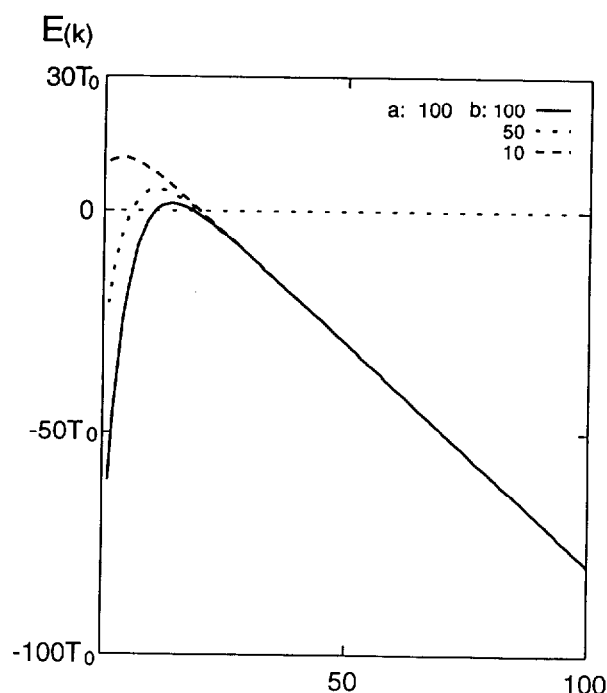
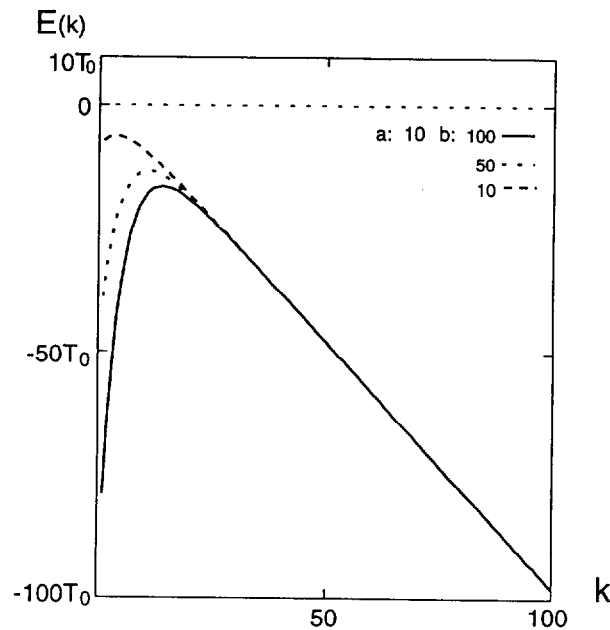


図 7.8: 探索利得 E と探索回数 k の関係 その 5

かわりなく $E(k)$ は正の値をとる。これは探索のどの場面においても探索は継続すべきであり、最後のサーバまで探索してみる価値があることを示している。 k の値が 100 に近付くと、 $E(k)$ の値は b の値にかかわらず全て $100T_0$ に収束する。これは、ペナルティが 0 に近付くことに加え、報酬が $1000T_0$ であるのに対して、最後に探索されるサーバにおいてもターゲットの存在確率が 0.2 であるためである。

図 7.8 では、探索の報酬を規定する a の値が図 7.7 と比較して小さいため、全体に $E(k)$ の値は小さくなる。ここで、ペナルティを規定する定数 b が最小である 10 をとる場合は、 $k = 20$ までは $E(k)$ の値は正であるが、これを境界として負へ転じ、単調減少する。これは、 k が 20 を越えてもターゲットが発見できない場合は探索を打ち切るべきであることを示している。

b が 50 または 100 の場合には、 $E(k)$ の値は負からはじまり、 k がある値をとる間は正に転じ（例えば、 $b = 50$ の場合は k が 6 から 19 の間）、さらにその後、負へ戻り、単調に減少する。このような特性を示す場合、探索の方針は 2 つ考えられる。まず、何らかの理由で探索回数 k を大きくできず、 $E(k)$ の値が負から正に転ずる前に探索を打ち切らなければならない場合は、1 回も探索を行うべきではない。次に、それ以上の探索が可能であれば、少なくとも $E(k)$ が正から負に転ずるまでの間は探索を継続すべきである。また、この場

図 7.9: 探索利得 E と探索回数 k の関係 その 6

合でもそれ以上の回数の探索はすべきではない。

図 7.9 では、探索の報酬を規定する a の値がさらに小さくなり、ペナルティを規定する定数 b の大きさによらず全ての場合で探索による利得が得られる期待は薄い。図 7.5 の場合と同様に $a = 10$ であるから、特に k が 10 を越える場合の探索にメリットは無い。これらの場合も b の値にかかわらず全て探索を 1 回も行うべきではない。

7.5 まとめ

本章では、超分散環境に散在する情報の効率的な探索方式を考えた際、特にノード数が多く、探索の打ち切りを考慮するのが現実的である場合の探索方式について検討を行った。

具体的には、探索エラーが生じる危険率を一定値以下に抑制する探索回数を計算し、この回数を最大探索回数として探索打ち切りの基準とする手法、および、探索報酬と探索エラーによるペナルティを定義し、探索から得られる利得の期待値を見積もることによって探索者にとって有利な回数まで探索を継続する手法を提案した。

今後はさらに提案方式の有効性について詳細に検討し、実際のインターネットに適用できる手法を模索していく。また、7.4 で提案した手法の探索報酬およびペナルティについて、具体的な適用例を探っていく予定である。

第 8 章

結論

昨今、World Wide Web(以下 WWW と略)のような情報空間が、我々の生活に密接に結び付いたものとして定着するにつれ、WWW のもつ情報量、あるいはそれらの情報へのアクセス数の増加は著しく大きくなってきている。

WWW のような、情報を一元的に管理しない巨大な情報空間では、ネットワークユーザが必要とする情報がネットワーク内に既に発信されているにもかかわらず、それを発見できないという問題が生じる。また、必要な情報を得るために、ネットワーク内を無作為に探索することは、探索時間と探索コストにロスを生む。そこで、このような環境における効果的な情報探索法について論じることはこれからますます重要になるものと考えられる。

そこで、本研究では、大規模ネットワーク上に散在する情報について、特に画像情報のような大容量を要するデータを中心に、効率的な探索方式を検討した。

このような探索問題は、単一標的問題と、不定標的問題の 2 つに分けて考えられるが、両者について考察している。

単一標的問題については、探索情報の存在位置に関する曖昧さをエントロピーで表し、このエントロピーの減少量が最大となるような探索方針を常に選択していくという手法を提案している。また、この手法を用いた画像情報探索システムのプロトタイプを設計・製作し、本研究の有効性を確認している。

また、不定標的問題については、探索に要する通信・処理コストと時間コストを定義し、これら両方について、探索終了までの平均コストの予測値を求め、それらの荷重和が最小となる方式を提案し、他の方法と比較して提案方式の有効性を確認している。また、エージェントを用いた探索方式について、分類し、それらのいくつかについて、効率的な探索方式を提案し、有効性を確認した。さらに、多数のノードの探索に伴う、探索打ち切りの指針についても示す方式を提案した。

以上の検討は、超分散データベース問題の一部に過ぎず、これからこの分野のさらなる

研究が期待される。

本研究において、最も困難な問題は、各ノードにおけるターゲットの存在確率の正確な推定方法である。これまでは、情報の類似性、および過去の探索経験を用い、推定する方法をとってきた。しかし、情報同士の類似性については、意味的に求めることは大変困難であり、現在のところ、一部の分野に限定された探索のみ有効な場合があると思われる。画像情報の特徴量から客観的な類似性を求めることはある程度可能であると思われるが、意味的に限定せずに膨大な画像情報全てを探索対象におくことも大変困難である。

一方、過去の探索経験を用いることは大変有効な手段であると思われるが、全く探索経験の無い状態では効果は無い。よって、探索経験が無い場合でも初期値として何らかの経験に相当するデータを持たせ、経験を積むことにより、各探索者に固有の探索用知識を持たせられるようにすることが考えられる。しかし、探索者の嗜好により、得意な分野とそうでない分野が生じ、不得意な分野では探索の精度が落ちることは容易に推測できる。

この部分の追求は本研究では残念ながら不十分に終わったが、今後の課題として、非常に重要であると思われる。

謝辞

本研究を進めるに当たり，終始辛抱強く御指導下さった酒井善則教授に感謝致します。また，常に的確なアドバイスを与えて下さった吉田俊之助教授に感謝致します。さらに研究上の有益な御意見・御助言を頂きました山岡克式助手，並びに酒井研究室の皆さんに感謝致します。

参考文献

- [1] 菊井, 鷺崎, 林, 砂場: “インターネット情報ナビゲーションにおける多言語機能” 情報学 自然言語処理の応用に関するシンポジウム (1995)
- [2] 菊井, 鷺崎, 林, 砂場: “WWW 情報空間における Resource Discovery と Navigation 支援” 信学技報 AI95-31, pp.71-78 (1995-09)
- [3] アリ・ウイドド, 長谷川: “出張情報フィルタリングエージェント” 信学技報 IN97-55, pp.57-62 (1997-06)
- [4] 國頭, 奥村, 相澤, 羽鳥: “協調エージェント間通信のための Tracking Agent に関する検討” 信学技報 IN97-20, CS97-1, MVE97-1, pp.1-8 (1997-04)
- [5] goo, <http://www.goo.ne.jp/>
- [6] Open Text Index, <http://www.jp.opentext.com/>
- [7] Lycos home page, <http://www.lycos.com/>
- [8] Infoseek Japan, <http://www.infoseek.co.jp/>
- [9] ingrid, <http://www.ingrid.org/>
- [10] Alta Vista, <http://www.altavista.digital.com/>
- [11] Hole-in-One, <http://hole-in-one.com//>
- [12] Image Surfer, <http://isurf.interpix.com//>
- [13] 西村, 河野, 長谷川: “WWW データ資源探索システムの実装と評価”, 信学技報 DE96-54 pp.1-6 (1996-07)

- [14] 高羽, 谷田部, 佐藤, 坂内: “World Wide Web 上の公開型画像検索システム GIRLS”, 信学技報 IE95-76, pp.1-8 (1995-11)
- [15] 鳥居: “意味解析機能を備えた WWW 検索システム”, 信学技報 DE96-56, pp.13-18 (1996-07)
- [16] 西尾, 岸本: “エージェントの負荷を考慮した分散協調による複数回線設定”, 信学技報 CS94-81, pp.49-54 (1994-08)
- [17] Peter Ingwersen 原著, 藤原 鎮男 監訳, 細野 他訳: “情報検索研究-認知的アプローチ-”, トッパン (1995)
- [18] David Ellis 原著, 細野 公男 監訳, 齊藤 他訳: “情報検索論-認知的アプローチへの展望-”, 丸善 (1994)
- [19] 宮川 洋, 原島 博, 今井 秀樹: “情報と符号の理論”, 岩波書店 (1983)
- [20] 柴田, 井上: “画像データベースの連想検索方式”, 電子情報通信学会論文誌, Vol.J73-D-II, No.4, pp.526-534 (1990)
- [21] 加藤, 栗田: “画像の内容検索-電子美術館への応用-”, 情報処理 33, 5, pp.466-477 (1992).
- [22] 栗田, 加藤: “印象語による絵画データベースの検索”, 情報処理学会論文誌 Vol.33, No.11, pp.1372-1383 (1992)
- [23] 村松, 酒井: “構造化情報を利用した画像検索法”, テレビ誌 47, 9, pp.1229-1236 (1993).
- [24] 岡野, 木下, 酒井: “手書き画像とのマッチングによる画像検索”, 信学技報 IE93-96, pp.53-60 (1994-01)
- [25] Myron Flickner et al.: “Query by Image and Video Content: The QBIC System”, Computer, 1995 September, pp.23-31
- [26] James Hafner et al.: “Efficient Color Histogram Indexing for Quadratic Form Distance Functions”, IEEE Pattern Analysis, Vol.17, No.7, July pp.729-736, 1995
- [27] Goh, 吉田, 酒井: “色情報のヒストグラムを用いた画像検索法”, 1995 信学ソ大, D-170
- [28] 国立国語研究所: “分類語彙表”, 秀英出版 (昭和 39 年)

-
- [29] 平岡 裕一, 山岡克式, 吉田 俊之, 酒井 善則: “画像探索のためのマルチエージェントの構築” 信学技報 IN97-174, pp.97-104 (1998-02)
- [30] 石畑 清: “アルゴリズムとデータ構造”, 岩波書店 (1989)
- [31] 多田和夫: “探索理論” 日科技連出版社 (1973)

付録 A

5.3 節の定理の証明

5.3 節の定理を，1 箇所ずつの探索が最適であること（証明 1），ノードは c_i/P_i の昇順に探索するのが最適であること（証明 2），の 2 段階にわけて証明する．

<証明 1 >

ノード総数を $N(\geq 2)$ として，第 i ノードを探索するための通信コストが c_i ，第 i ノードにおけるターゲットの存在確率が P_i であると仮定する．簡単のため， i の昇順に従って探索を行うものとし，全てのノードを 1 箇所ずつ探索した場合の期待通信コストの和を ε_1 ， m 回目 ($m < N$) の探索だけ，第 m ノードと第 $m+1$ ノードを同時に探索した場合の期待通信コストの和を ε_2 とする．

従って，

$$\begin{aligned}\varepsilon_1 = & c_1 P_1 + (c_1 + c_2)(1 - P_1)P_2 + \\ & (c_1 + c_2 + c_3)(1 - P_1)(1 - P_2)P_3 + \cdots \\ & \cdots + (c_1 + c_2 + c_3 + \cdots + c_{m-1}) \\ & \times (1 - P_1)(1 - P_2) \cdots (1 - P_{m-2})P_{m-1} \\ & + (c_1 + c_2 + c_3 + \cdots + c_{m-1} + c_m) \\ & \times (1 - P_1)(1 - P_2) \cdots (1 - P_{m-1})P_m \\ & + (c_1 + c_2 + c_3 + \cdots + c_{m-1} + c_m + c_{m+1}) \\ & \times (1 - P_1)(1 - P_2) \cdots (1 - P_m)P_{m+1} + \cdots \\ & \cdots + (c_1 + c_2 + c_3 + \cdots + c_N) \\ & \times (1 - P_1)(1 - P_2) \cdots (1 - P_{N-1})P_N \\ & \cdots + (c_1 + c_2 + c_3 + \cdots + c_N) \\ & \times (1 - P_1)(1 - P_2) \cdots\end{aligned}$$

$$(1 - P_{N-1})(1 - P_N) \quad (\text{A.1})$$

$$\begin{aligned}
\varepsilon_2 = & c_1 P_1 + (c_1 + c_2)(1 - P_1)P_2 + \\
& (c_1 + c_2 + c_3)(1 - P_1)(1 - P_2)P_3 + \cdots \\
& \cdots + (c_1 + c_2 + c_3 + \cdots + c_{m-1}) \\
& \quad \times (1 - P_1)(1 - P_2) \cdots (1 - P_{m-2})P_{m-1} \\
& + (c_1 + c_2 + c_3 + \cdots + c_{m-1} + c_m + c_{m+1}) \\
& \quad (1 - P_1)(1 - P_2) \cdots (1 - P_{m-1}) \\
& \quad \times \{1 - (1 - P_m)(1 - P_{m+1})\} + \cdots \\
& \cdots + (c_1 + c_2 + c_3 + \cdots + c_N) \\
& \quad \times (1 - P_1)(1 - P_2) \cdots (1 - P_{N-1})P_N \\
& \cdots + (c_1 + c_2 + c_3 + \cdots + c_N) \\
& \quad \times (1 - P_1)(1 - P_2) \cdots \\
& \quad (1 - P_{N-1})(1 - P_N) \quad (\text{A.2})
\end{aligned}$$

簡単のため,

$$c_1 + c_2 + c_3 + \cdots + c_{m-1} = \mathbf{C} \quad (\text{A.3})$$

$$(1 - P_1)(1 - P_2) \cdots (1 - P_{m-1}) = \mathbf{P} \quad (\text{A.4})$$

とおくと, $\mathbf{P} > 0$ であるから,

$$\begin{aligned}
\varepsilon_2 - \varepsilon_1 = & (\mathbf{C} + c_m + c_{m+1}) \\
& \quad \times \mathbf{P}\{1 - (1 - P_m)(1 - P_{m+1})\} \\
& \quad - (\mathbf{C} + c_m)\mathbf{P}P_m - (\mathbf{C} + c_m + c_{m+1}) \\
& \quad \quad \times \mathbf{P}(1 - P_m)P_{m+1} \\
= & \mathbf{P}\{(\mathbf{C} + c_m + c_{m+1}) \\
& \quad \times (P_m + P_{m+1} - P_m P_{m+1}) \\
& \quad - (\mathbf{C} + c_m)P_m - (\mathbf{C} + c_m + c_{m+1}) \\
& \quad \quad \times (P_{m+1} - P_m P_{m+1})\} \\
= & c_{m+1}\mathbf{P}P_m \quad (\text{A.5})
\end{aligned}$$

従って1箇所ずつ探索した場合の方が、2箇所まとめて探索した場合よりも期待通信コストは小さい。

このことは、どのような組み合わせでまとめた探索よりも、その部分を別々の回に分けて探索した方が期待通信コストが小さくできることを意味する。すなわち通信コストのみを考えた場合、1箇所ずつの探索を行うのが最善である。

<証明 1 終わり>

<証明 2>

証明 1 と同様に、 i の順番に従って1箇所ずつ探索を行った場合の期待コストは ε_1 である。 m 番めと $m+1$ 番めのノードの探索順を入れ換えたときの期待コストを ε_3 とすると、

$$\begin{aligned}
\varepsilon_3 = & c_1 P_1 + (c_1 + c_2)(1 - P_1)P_2 + \cdots \\
& \cdots + (c_1 + c_2 + \cdots + c_{m-1} + c_{m+1}) \\
& \quad \times (1 - P_1)(1 - P_2) \cdots (1 - P_{m-1})P_{m+1} \\
& + (c_1 + c_2 + \cdots + c_{m-1} + c_{m+1} + c_m) \\
& \quad \times (1 - P_1)(1 - P_2) \cdots (1 - P_{m-1}) \\
& \quad \times (1 - P_{m+1})P_m + \cdots \\
& \cdots + (c_1 + c_2 + \cdots + c_N) \\
& \quad \times (1 - P_1)(1 - P_2) \cdots (1 - P_{N-1})P_N \\
& \cdots + (c_1 + c_2 + \cdots + c_N) \\
& \quad \times (1 - P_1)(1 - P_2) \cdots \\
& \quad (1 - P_{N-1})(1 - P_N)
\end{aligned} \tag{A.6}$$

ここでも証明 1 と同様に、 \mathbf{C} , \mathbf{P} を用いると、

$$\begin{aligned}
\varepsilon_1 - \varepsilon_3 = & (\mathbf{C} + c_m)\mathbf{P}P_m \\
& + (\mathbf{C} + c_m + c_{m+1})\mathbf{P}(1 - P_m)P_{m+1} \\
& \quad - (\mathbf{C} + c_{m+1})\mathbf{P}P_{m+1} \\
& - (\mathbf{C} + c_{m+1} + c_m)\mathbf{P}(1 - P_{m+1})P_m \\
= & \mathbf{P}(P_{m+1} \cdot c_m - P_m \cdot c_{m+1})
\end{aligned} \tag{A.7}$$

$$\frac{c_m}{P_m} \geq \frac{c_{m+1}}{P_{m+1}} \tag{A.8}$$

とすると、 $P > 0$ より、

$$\varepsilon_1 - \varepsilon_3 \geq 0 \quad (\text{A.9})$$

よって、連続する2回の探索では、 c_i/P_i が小さいノードを先に行う場合期待コストは小さくなる。このことから最小の期待コストを得るには、探索順を c_i/P_i の昇順とすればよいことは明らかである。

<証明 2 終わり>

付録 B

$E(n_j + n_k) > E(n_k)$ の証明

<証明>

$$E(n_j + n_k) = \frac{c_j + c_k}{1 - (1 - P_j)(1 - P_k)} \quad (\text{B.1})$$

また、第 k ノードのみの探索は、

$$E(n_k) = \frac{c_k}{P_k} \quad (\text{B.2})$$

よって、

$$E(n_j + n_k) - E(n_k)$$

$$\begin{aligned} &= \frac{c_j + c_k}{1 - (1 - P_j)(1 - P_k)} - \frac{c_k}{P_k} \\ &= \frac{P_k(c_j + c_k) - (P_j + P_k - P_j P_k)c_k}{(P_j + P_k - P_j P_k)P_k} \\ &= \frac{(c_j P_k - c_k P_j) + c_k P_j P_k}{(P_j + P_k - P_j P_k)P_k} \end{aligned} \quad (\text{B.3})$$

とすると、 $(P_j + P_k - P_j P_k)P_k > 0$ と (5.5) 式より、

$$E(n_j + n_k) > E(n_k) \quad (\text{B.4})$$

となる。

<証明終わり>

付録 C

$E(n_1 + n_2 + \cdots + n_n) > E(n_n)$ の証明

<証明>

(i) n_1 と n_2 について, (B.4) 式より

$$E(n_1 + n_2) > E(n_2) \quad (\text{C.1})$$

は明らか.

(ii) $m(\geq 3)$ 個のノード n_1, n_2, \cdots, n_m について,

$$E(n_1 + n_2 + \cdots + n_m) > E(n_m) \quad (\text{C.2})$$

が成立すると仮定すると,

$$\begin{aligned} & E(n_1 + n_2 + \cdots + n_m + n_{m+1}) \\ &= \frac{c_1 + c_2 + \cdots + c_m + c_{m+1}}{1 - (1 - P_1)(1 - P_2) \cdots (1 - P_m)(1 - P_{m+1})} \\ &= \frac{c_1 + c_2 + \cdots + c_m + c_{m+1}}{1 - [1 - \{1 - (1 - P_1) \cdots (1 - P_m)\}](1 - P_{m+1})} \end{aligned}$$

ここで,

$$c_1 + c_2 + \cdots + c_m = C_{uni} \quad (\text{C.3})$$

$$1 - (1 - P_1)(1 - P_2) \cdots (1 - P_m) = P_{uni} \quad (\text{C.4})$$

とおくと,

$$\begin{aligned} & E(n_1 + n_2 + \cdots + n_m + n_{m+1}) \\ &= \frac{C_{uni} + c_{m+1}}{1 - (1 - P_{uni})(1 - P_{m+1})} \end{aligned} \quad (\text{C.5})$$

ところで, C_{uni} と P_{uni} は最初の n_1 から n_m までのノードをまとめてひとつのノード (\mathbf{n}_{uni} とする) として見た場合の通信コストとターゲットの存在確率にほかならない. (C.2) 式の仮定より,

$$\begin{aligned} \frac{C_{uni}}{P_{uni}} &= E(n_1 + n_2 + \cdots + n_m) > E(n_m) = \frac{c_m}{p_m} \\ &\geq \frac{c_{m+1}}{p_{m+1}} \end{aligned} \quad (C.6)$$

よって, (B.4) 式より,

$$\begin{aligned} E(n_1 + n_2 + \cdots + n_m + n_{m+1}) &= E(\mathbf{n}_{uni} + n_{m+1}) \\ &> E(n_{m+1}) \end{aligned} \quad (C.7)$$

(iii) (i),(ii) のことから数学的帰納法より,

$$E(n_1 + n_2 + \cdots + n_n) > E(n_n) \quad (C.8)$$

が示された.

<証明終わり>

目次

2.1	超分散データベースモデル	4
4.1	平均探索コスト	21
4.2	平均探索時間	22
4.3	システム概念図	24
4.4	類似度グラフとハイパーリンク	25
4.5	MMI の表示画面例	27
5.1	離散確率分布	34
5.2	重み比と通信コスト	34
5.3	重み比と時間コスト	35
5.4	重み比と総コスト	35
5.5	誤差を含む離散確率分布	36
5.6	確率分布に誤差を含む場合の総コスト	38
5.7	方式 A および B を用いた総コストの比較	39
6.1	単体・機動性なしのエージェントの探索	47
6.2	単体・機動性ありのエージェントの探索	51
6.3	離散確率分布	53
6.4	Single-Mobile Type 総探索コスト ($W_1 = 1.0, W_2 = 0$)	54
6.5	Single-Mobile Type 総探索コスト ($W_1 = 0.5, W_2 = 0.5$)	54
6.6	複数・機動性なしのエージェントの探索	56
6.7	エージェントの配置	59
6.8	Multi-Fixed Type 総探索コスト ($W_1 = 1.0, W_2 = 0$)	60
6.9	Multi-Fixed Type 総探索コスト ($W_1 = 0.5, W_2 = 0.5$)	60
7.1	システムイメージ	63

7.2	離散確率分布 その1	66
7.3	探索利得 E と探索回数 k の関係 その1	67
7.4	探索利得 E と探索回数 k の関係 その2	68
7.5	探索利得 E と探索回数 k の関係 その3	68
7.6	離散確率分布 その2	69
7.7	探索利得 E と探索回数 k の関係 その4	69
7.8	探索利得 E と探索回数 k の関係 その5	70
7.9	探索利得 E と探索回数 k の関係 その6	71

表 目 次

6.1 エージェントの分類	45
---------------------	----

本研究に関する発表論文

研究論文

- 酒井 善則, 山岡 克式, 菅原 真司: “エントロピーに基づいた情報探索に関する検討” (ショートノート), 信学論 (A), Vol.J78-A No.7 pp.894-896, 1995
- 菅原 真司, ゴー ホックチュー, 大槻一博, 酒井 善則: “分散環境における画像情報探索システム”, 画電誌 第 26 卷 第 4 号 (1997) (画像電子学会論文賞受賞)
- 菅原 真司, 山岡 克式, 酒井 善則: “ネットワークにおける画像情報の効率的探索法に関する検討”, 信学論 B-I Vol. J81-B-I No.8 pp.484-493 (1998 年 8 月)
- 菅原 真司, 大島 貴光, 山岡 克式, 酒井 善則: “エージェントを用いたネットワークにおける情報探索に関する検討”, 信学論 B 採録決定

国際会議

- S.SUGAWARA, K.YAMAOKA, Y.SAKAI: “A Study on Image Searching Method in Super Distributed Database”, IEEE GLOBECOM '97, pp.736-740, 1997

国内学会発表 (研究会)

- 菅原真司, 大槻 一博, 山岡 克式, 酒井 善則: “分散データベースにおける画像情報探索システム”, 信学技報 IN96-39 pp55-62 (1996-07) (1996 年 7 月 大阪 阪大)
- 山岡 克式, 菅原 真司, 大槻 一博, 酒井 善則: “エントロピーに基づいた分散画像情報探索システム”, 画電学会第 152 回研究会講演予稿 96-02-9 pp33-36 (1996 年 9 月 東京 早大)
- 菅原 真司, 山岡 克式, 酒井 善則: “超分散データベースにおける情報探索のためのルーチング方式”, 信学技報 IN96-112, OFS96-50 pp23-30 (1997-01) (1997 年 1 月 京都 ATR)

- 菅原 真司, 山根 裕明, 酒井 善則：“超分散環境における効率的情報探索法”, 信学技報 SSE97-116, IN97-109, CS97-107 pp67-72 (1997-09) (1997年9月 仙台 東北大)
- 菅原 真司, 山岡 克式, 酒井 善則：“インターネットにおけるブラウザ検索効率の最適化について”, 信学技報 SSE98-92, IN98-73, CS98-89 pp19-24 (1998-09) (1998年9月 仙台 東北大)
- 菅原 真司, 山岡 克式, 酒井 善則：“超分散情報探索手法に対する探索打ち切りの効果”, 信学技報 IN98-103, CQ98-45 pp1-6 (1998-10) (1998年10月 東京 機械振興会館)

国内学会発表（電子情報通信学会 全国大会, 総合大会, ソサイエティ大会）

- 菅原真司, 山岡克式, 酒井善則：“複数最小木によるメディア探索ネットワークの設計” 信学春季全大 B-684 (1994年3月 横浜 慶大)
- 菅原真司, 山岡克式, 酒井善則：“独立した複数データベースの検索方式” 信学秋季全大 B-674 p272 (1994年9月 東京 中央大)
- 酒井善則, 山岡克式, 菅原真司：“エントロピーに基づいた情報探索法” 信学総大 B-831 p277 (1995年3月 福岡 福岡工業大)
- 菅原真司, 山岡克式, 酒井善則：“複数最小木を用いた階層型 VOD 設計法” 信学総大 SB-4-1 pp.639-640 (1995年3月 福岡 福岡工業大)
- 菅原真司, 大槻一博, 山岡克式, 吉田俊之：“超分散データベースネットワークの構成法” 信学春大 B-823 p255 (1996年3月 東京 東工大)
- 小室智之, 菅原真司, 酒井善則：“超分散データベースネットワークにおける情報探索方式” 信学春大 B-824 p256 (1996年3月 東京 東工大)
- 菅原真司, 山岡克式, 酒井善則：“超分散データベース探索に関する一検討” 信学総大 B-7-191 p320 (1997年3月 大阪 関西大)
- 菅原真司, 山岡克式, 酒井善則：“ネットワークにおける画像探索の効率化に関する検討” 信学総大 B-7-159 p280 (1998年3月 平塚 東海大)

- 菅原真司, 山岡克式, 酒井善則: “探索打ち切りを考慮したネットワークにおける効率的な情報探索法” 信学ソ大 B-7-7 p128 (1998年9月 甲府 山梨大)