

情報単位 Capta を用いた情報流通システム RRR の設計と実装\*

6 X - 1

宮川 祥子†

慶應義塾大学 政策・メディア研究科‡

清木 康§

慶應義塾大学 環境情報学部¶

金子郁容||

慶應義塾大学 政策・メディア研究科‡

1 はじめに

WWW に代表される情報システムの利用は、おおむね (1) 情報提供者が作成したリンクをトレースする、(2) サーチエンジンを利用してキーワードにマッチする URL リストを得る、という2つの方法に限られていた。インターネット上の情報システムは、多くのユーザが同一のサーバにアクセスするという特徴を持つにもかかわらず、サーバに蓄積されたユーザの情報はシステムにもユーザにも活用されず、利用効率の改善に活かされてこなかった。また、ユーザ個人にとっても、自分が得たノウハウを蓄積し再利用することは難しかった。最近では、利用者の嗜好を分析し、好みにあった情報を紹介するシステムがいくつか開発されているが [1]、検索の過程で得られたノウハウがユーザ間で十分に共有されているとは言いがたい。本稿では、ユーザが情報検索プロセスの中で得られたノウハウをユーザ間で共有、再利用するための対象データ Capta の設定と、その処理系である RRR(Reflektiv Recommendation Retrieval) システムについて述べる。

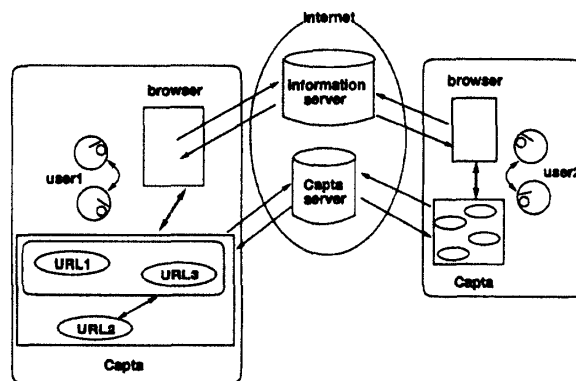


図 1: Capta 交換/共有モデル

- 記録した URL、リンクに対してのコメントの設定をサポートする

クライアントで作られた Capta は、ネットワーク上の Capta サーバに送られ、蓄積される。ユーザは、蓄積された Capta を検索することで、同じ目的で他のユーザによって作られた Capta を参照することができる (図 1)。

2 capta 交換/共有方式の概要

Capta とは、アメリカの心理学者レインによる造語で、「意味を持つ情報単位」を指している。本稿では、Capta に基づく情報システム研究の第一段階として、情報検索の過程で得られた情報の場所、情報どうしを関連づけるリンク、補足的に書かれるコメントを編集したメタ情報を Capta とする。Capta を編集するためのツールは Capta エディタと呼ばれ、以下のような機能を持つ。

- 検索で見つかった役に立つ情報の URL を記録する
- 記録した URL をグループ化したり関連づけるためのリンクを提供する

3 Capta の対象データ

Capta は以下のメタデータとして構成される (図 2)。

- URL
- URL の示すコンテンツのインデックス
- オブジェクト (URL あるいはあるいは URL 群をまとめた単位)
- オブジェクト識別子
- オブジェクト間の抽象的な関係を表す関係識別子
- 関係識別子に対してつけられる名前

コンテンツのインデックスとは、コンテンツにどのような単語が含まれるかを表すデータ構造である。インデックスは Capta 編集時に自動的に作成され、Capta を検索する際に用いられる。

オブジェクト間の抽象的な関係性とは、「上位-下位」「具体化-抽象化」「対立」「原因-結果」「論理的帰結」等の関係を明示的に表すデータであり、オブジェクト間の関連性の記述に用いられる。それぞれの関係性には具体的に名前がつけられる。

\*Design and Implementation of RRR—the Information System Based on Information Unit "Capta"

†Shoko Miyagawa

‡Graduate School of Media and Governance, Keio University

§Yasushi Kiyoki

¶Department of Environmental Information, Keio University

||Ikuyo Kaneko

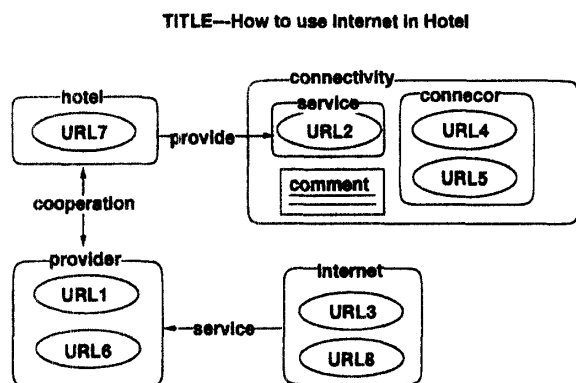


図 2: Capta の例

## 4 Capta インデックスを対象とした検索

ユーザによって編集された Capta は、ネットワーク上の Capta サーバに送られ、Capta DB に蓄積される。Capta の検索は、URL インデックスを対象とした複数キーワードの条件式 (and/or の組合せ) で行なわれる。Capta サーバは、問い合わせのキーワード群に対して Capta のインデックスを検索し、and の場合は全てのキーワードを含む Capta を、or の場合はいずれかのキーワードを含む Capta を抽出する。

## 5 実装

RRR は、WWW 上に実装された Capta 共有システムである。RRR では、既存の WWW サーバやクライアントを変更することなく Capta 共有機構を実現するために、Capta サーバとして WWW の Proxy サーバを用いた。Proxy-Capta サーバは、WWW サーバから受け取ったデータをクライアントに渡す際にクライアント側の Capta エディタへのインタフェースを付加する。このインタフェースは、WWW クライアントに送られたデータの URL とコンテンツを Capta エディタに渡す。Capta エディタは、Capta の編集、Proxy-Capta サーバとの通信を行なう (図 3)。

## 6 Capta 検索による情報抽出の効果

RRR ユーザは、Capta を相互に共有し、分散ネットワーク上に蓄積された Capta を検索することにより、コンテキストに沿って編集された情報の集合を得ることが可能となる。また、コンテキストベースの検索が可能になり、曖昧な問い合わせに対しても有効な検索結果を得ることができる。例えば、旅先でインターネットに接続できるかどうかを知りたいユーザがいた場合、

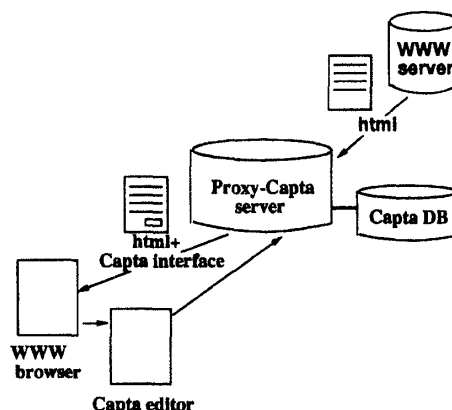


図 3: RRR システム構成

従来の検索機構では、サーチエンジンを使って「インターネット & ホテル」というキーワードで検索をしても、ある HTML 文書に両方の語が記述されていないとマッチしない。リンクをトレースする場合でも、「ホテル」のリンク集がインターネットが利用可能かどうかで分類されているとは限らない。Capta を利用した検索の場合、ホテルについての情報とインターネットの繋ぎ方についての情報が別々の HTML 文書にあっても、それらの HTML 文書の URL が同一の Capta に含まれていれば Capta 単位では「インターネット & ホテル」にマッチした情報を抽出することが可能となる。また、ある Capta と関連性のある Capta を検索することで、コンテキストの意味空間を生成できるので、「ホテル」で検索した場合でも「旅館」の情報も検索することができるようになる。このように、Capta による情報検索は、一般的には関連の少ないと思われるキーワード群による検索や、キーワードの意味を一般化した検索に対して効果があることがわかる。また、同じ興味を持つユーザどうしが Capta を共有することで、ネットワーク上に共同知を創出することが可能となる。

## 7 おわりに

本稿では、情報検索で得られた知識を表現する Capta の設定と、Capta を用いた情報流通システム RRR の実装方式について述べた。今後の課題は、情報編集に適した Capta データモデルの設計、データマイニングの手法を用いた Capta 間の関連性抽出である。

## 参考文献

- [1] Paul Resnick, Hal R. Varian *et. al.*: Recommender Systems, CACM, Vol.40, No.2,1997