知能と複雑系　132－7
(2003. 3. 13)

時間を考慮した知的情報アクセス技術獲得システム

後藤　文太朗†

†北見工業大学　情報システム工学科

〒090-8507　北見市公園町165番地

E-mail:　† fg@cs.kitami-it.ac.jp

あらまし　SPIRALは，Webを利用した業務及び知的生産作業の支援のための知的情報統合利用環境である．SPIRALでは，情報の収集，編集，発信（共有）を連続的に繰り返し行っていくというワークフローを取り込んでいる．収集に関して，ユーザが見たWebページのURLだけではなく，その内容もすべてアクセス時刻と対応付けて保存し，参照できる機能を持つ．我々は，この時間情報が付加されたデータを用いた知的情報アクセス技術を獲得するためのシステムを提案してきた．本稿では，大規模な実験を行う際のパフォーマンスの問題，時間を考慮した複数アクセスデータからの訓練事例作成に関しての対応について報告する．

キーワード　データマイニング，時間，アクセス技術，情報統合

Intelligent-Information-Access-Method Acquisition System using Temporal Information

Fumitaro GOTO †

† Department of Computer Sciences, Kitami Institute of Technology

165, Koen-cho, Kitami, 090-8507 Japan

E-mail:　† fg@cs.kitami-it.ac.jp

Abstract　WWW and e-mail are currently used for various purposes. To achieve these various purposes, we use all sorts of tools. But, there exists a great gap between those true purposes and works that use those tools. WWW and existing tools force us to fill the gap. So, we have proposed the SPIRAL system, which is a seamless user environment for WWW. In SPIRAL, when user access to web sites, SPIRAL stores not only the URLs but also the contents with time stamp. So, we have proposed information-access-method acquisition system using these temporal information. In this article, we report how to solve performance problems when we use a large scale of data and how to create a training example from multiple access data.

Keyword　Data Mining, Temporal Information, Access Method, Information Integration

1. はじめに

現在，インターネット上では，WWW(World Wide Web)や電子メールが多様な目的で使われている．それらの目的の種別は，商品案内・販売，専門知識の公開といったものから，個人的な連絡にいたるまで非常に広いものになっている．今後，社会生活におけるインターネットの重要性はさらに増加していくことは間違いないであろう．

これらの多様な目的を達成するために，我々は各種ツールを利用する．膨大なWWW上の情報を利用するユーザを支援するものには，URL管理ツール，WWWサイトのオフラインブラウジングツール，検索サーバ，ポータルサイトといったものがある．また，LogicWeb[1]にみられるようなWebページ概念の拡張も行われてきている．WWW上に情報を公開するユーザを支援するものには，HTMLエディタや，リンクやドキュメントの管理機能を持ったWWWサイト管理ツールといったものがある．AppleのCypberdog[2]は，Webページをドキュメントのコンポーネントとすることを可能にしたものである．WWW上の情報の利用形態を広げている．

しかし，実際の目的と，これらのツールを使って行う作業との間には，依然，大きな隔たりがある．現在は，人間側がWWWや電子メール及び現状のツールに歩み寄ることにより，目的を達成しているといえる．

しかし，インターネットを使う機会，インターネット上の情報，インターネット上を飛び交う情報の増大により，「人

間側が歩み寄る」形でのインターネット利用のままでは，生産性はあがらなくなってしまう可能性が大きい．

この問題点を解決するために，我々はWWWのシームレスな利用環境として，SPIRAL[3]を提案した．

SPIRALを用いると，ユーザがアクセスしたWebサイトのURLだけではなく，そのコンテンツもタイムスタンプをつけてアクセスデータとしてデータベースに保存される．SPIRALへの機械学習機能モジュールの統合[7]，時間軸を考慮して複数アクセスデータを一つの訓練事例とした小規模な実験[10]を行ってきた．

これまでは，SPIRALはPrologを用いて実装が行われていた．しかし，この実装では，大規模な実験を行う際に，パフォーマンス上の問題があった．また，時間軸を考慮して複数アクセスデータを一つの訓練事例とする方法も限られていた．そこで，我々はそれらの点の解決を試みた．

まず，参考文献[3]で提案したSPIRALの概要を示し，次にSPIRALのアーキテクチャについて説明を行う．そして，現状の問題点について説明し，今回行った対応方法を説明する．

2. SPIRALの概要

SPIRALでは，Webサイトのダイナミック性と，WWW利用における複数ツールが独立しているために生じるデータ分散の問題点の解決を試み，既存ツールと比較して，人間側の歩み寄りを少なくした．

2.1. Webサイトのダイナミック性への対処

Webサイトは，その作成者によりダイナミックに更新されていく．これは，一般的には最新の情報を得られるということから利点といえるが，欠点もある．URLをもとに以前に訪れたWebサイトにアクセスしようとした時，その内容が変更されていたり，削除されていて，利用者の目的が達成できない場合がある．

URLのみを保存するようなブックマークでは，情報の保存形態としては不完全である．したがって，ブックマークを使っている限りは，その不完全さを人間が補って使うという，人間側からの歩み寄りが必要となる．

SPIRALでは，情報の保存形態として，URLだけではなく，Webページデータそのものを含むWWWにおけるアクセス活動全体の保存を行っている．これにより，先に述べたWebサイトのダイナミック性の問題点の解決を試みていた．

図1において，画面上部がSPIRALによって付加されたメニュー部で，下部がもともとのWebページである．上部のメニューにより，以前にアクセスしたWebページを参照することができる．

図1　SPIRAL利用画面1

2.2. データ分散への対処

ブラウジングはWebブラウザで行い，Webサイト構築はWebサイト構築ツールで行う．これら二つの作業で用いられるデータは，それぞれのツールで別々に扱われる．したがって，Webブラウザで見ているページにコメントをつけたり，複数のページをまとめたりして，それをWebページとして発信するといった作業をシームレスに行うことが困難である．

SPIRALでは，WWWの利用を単にブラウジングとしてではなく，情報の収集，分類，整理，編集，発信といった処理の繰り返しであるととらえた．各処理過程におけるデータを一つのオブジェクトデータベースに格納することで，データの共有化をはかった．そして，それらの処理がシームレスに螺旋を上に描いていくように行える利用環境を提供した．

図2は，SPIRALにおける編集・発信の例である．画面下部の左右それぞれが，ユーザがそれ以前にアクセスしたWebページであり，ユーザがその二つをグループ化して一つにし，WWW上に公開しているものである．

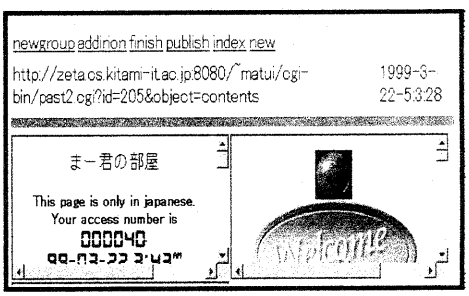

図2　SPIRAL利用画面2

3. SPIRALのアーキテクチャ

SPIRALのアーキテクチャは図3のようになり，以下のものから構成される．

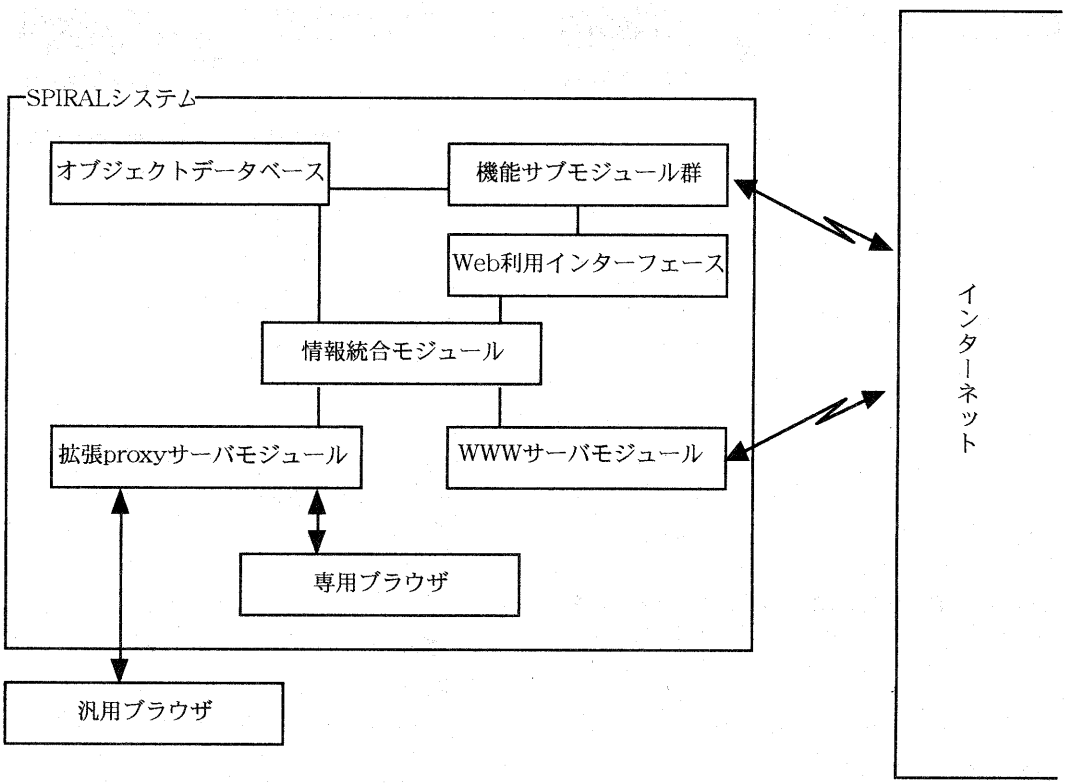

図3 SPIRALのアーキテクチャ

- オブジェクトデータベース
- 機能サブモジュール
- Web利用インターフェース
- 情報統合モジュール
- 拡張proxyサーバモジュール
- WWWサーバモジュール
- 専用ブラウザ

以下，これらの構成要素の機能について説明を行う．

3.1. オブジェクトデータベース

SPIRALで利用されるすべてのオブジェクトがオブジェクトデータベースに格納され，一元的に管理される．

ユーザがアクセスしたWebページデータ，それらのWebページをグループ化して新たに一つのWebページとしたデータ，外部公開したWebページデータが格納される．

SPIRALの支援範囲を，電子メール，ネットニュースといった他のインターネット上のメディアでの活動，オブジェクトデータベースの有効活用に関する活動まで拡張する機能を備えている．これらの活動は，次に述べる機能サブモジュール群により統一的に扱われ，それらの活動データがオブジェクトデータベースに格納される．

3.2. 機能サブモジュール群

支援活動の種別ごとに，その機能を実現する機構を，機能サブモジュールとして用意することで，アーキテクチャの一般化が行われている．

電子メールリーダー機能を実現した機能サブモジュールがある．POPにより電子メールを外部のサーバから取り出す機能，作成された電子メールを外部のSMTPサーバに送信する機能がここで実現されることになる．送受信されるデータは全て，オブジェクトデータベースに格納される．これにより，電子メールに関する活動の統合も可能となる．

編集の単位をWebページでから，より粒度の細かいものにすること及び編集機能を強化することも，それに対応した機能サブモジュールを用意することで実現される[6]．

オブジェクトデータベースの検索機能の強化に関しても，それに対応する機能サブモジュールを実装することにより行われる．現在は，Webに関するアクセス先の使用状況の検索が可能となっている[5]．

また，KDDツールであるC4.5[4]をオブジェクトデータベース中のデータに適用・利用する機能サブモジュールと

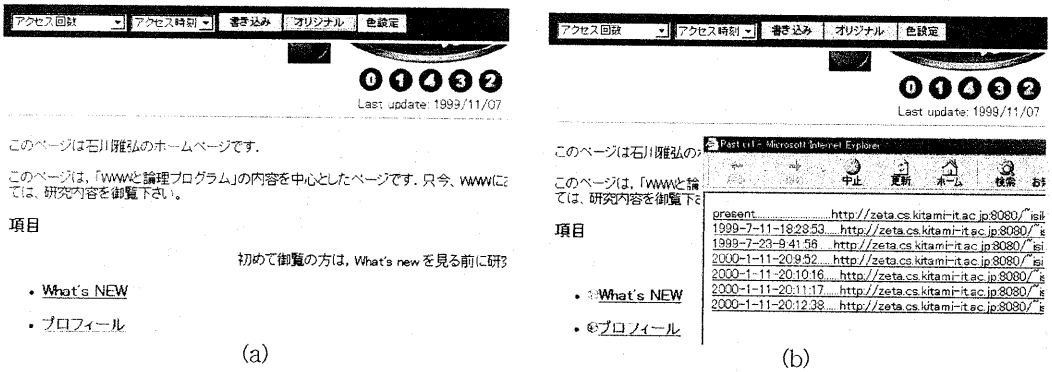

図4 情報統合の例

いったものもある[7]. オブジェクトデータベース中のどのオブジェクトを C4.5 に対するデータとして利用するかを指示する機能, オブジェクトデータベース中のオブジェクトを, C4.5 の訓練事例として適した属性 - 属性値の並びとクラスに変換することを実現する機能, 学習結果である決定木を新規のオブジェクトに対して適用する機能が個々で実現されることになる. 学習の際に使われたデータ, 学習結果のデータすべてがオブジェクトデータベースに格納される.

現在はまだ実装はされていないが, オブジェクトデータベースの圧縮等に関する機能, Web ページとして発信するデータの著作件に関するチェック機能といったことも, 機能サブモジュールとして実装していくことで, その対応が可能なアーキテクチャとなっている.

3.3. Web 利用インターフェース

前節で述べた機能サブモジュール群によって, 実現される機能は, Web 利用インターフェースにより SPIRAL 利用者に対する Web によるユーザインターフェースが与えられる.

Web により, 各機能を利用することで, それらの機能を利用した活動履歴が Web データとしてオブジェクトデータベースに保存される. Web データとしてそれらの活動履歴にアクセスできることで, それらをさらにデータとして編集作業を行っていくことが可能となる.

3.4. 情報統合モジュール

Web 利用インターフェースを介して利用している, 機能サブモジュールから得られるデータと, オブジェクトデータベース中にある他のデータとを情報統合するモジュールである.

このモジュールの利用例を図 4 に示す. システム実行時には, 図 4(a) の上部に示すようなシステム操作用のボタンが配置される. 書き込みボタンのクリックにより Web ページ上にイメージが埋め込まれ, 利用頻度等に応じて色分けされる. すなわち, 過去の利用頻度データ, 現在アクセスしている Web ページとが情報統合され, ユーザに見やすい形で提供される. 図 4(b) に示すように, イメージをなぞると実際のデータが表示され, イメージをクリックすると以前に使用した Web ページの内容へのリンクが利用できる.

3.5. 拡張 proxy サーバモジュール

拡張 proxy サーバモジュールは, Netscape Communicator や Internet Explorer 等の汎用的な Web ブラウザにおいて proxy サーバとして SPIRAL を設定することで, SPIRAL の機能をそれらのブラウザから利用できるようにするためのものである.

専用ブラウザにおいては, 編集機能の粒度をより細かくできるようにするために, 特化されたデータのやり取りを行う機能を実現している.

3.6. WWW サーバモジュール

WWW サーバモジュールは, オブジェクトデータベース中のオブジェクトをインターネット上に公開するための機能を実現する.

機能サブモジュール, Web 利用インターフェースの実装により, 旧来の Web ページだけではなく, 電子メール等の活動を, Web ページとして編集して発信することがシームレスに行うことができるようになっている.

3.7. 専用ブラウザ

編集単位の粒度をより細かなものとするために, 専用ブラウザのプロトタイプを Macintosh 上の IntelligentPad[8] システムを用いて開発を行った.

4. 大規模データへの対応

既存のシステムは, メインの実装言語が Prolog であったために, 実行パフォーマンスが悪いという問題点があった.

大規模データを収集して，実験を行う上で障害となっていた．

そこで，我々は開発環境としてWebObjects[11]を用い，メインの実装言語をJavaとすることで，この解決を試みた．

WebObjectsにおいては，Enterprise Objects Framework (EOF)が用意されており，データベーステーブルが Enterprise Objectsと呼ばれるJavaクラスのコレクションとして表現される．Enterprise Objectsでは，モデルと呼ばれる，データベースからオブジェクトへのマッピングが用意され，データベースとは独立にビジネスロジックをEnterprise Objectsクラスに記述することができる．また，Enterprise Objectsクラスのインスタンスは，「WebObjects」により分析され，データベース操作及び実行に関して，SQL文を直接記述する必要は基本的にない．これにより，データのやり取りがより効率的に行える仕組みが組み込まれている．

SPIRALにおけるオブジェクトデータベースをリレーショナルデータベースにより実装し，各モジュールに関するビジネスロジックを記述していくという対応をとることができる（図5）．これにより，Prologで行っていたよりも，高速なトランザクション処理が可能となり，加えて，WebObjectsによる無駄なデータベース処理の軽減という恩恵が得られる．現状では，拡張Proxyモジュールのプロトタイプが実装されている．

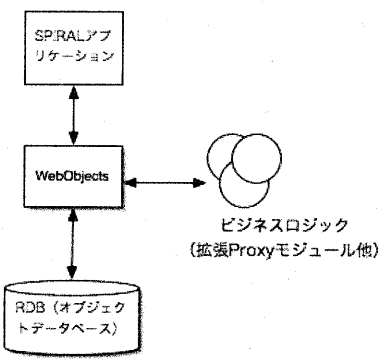

図5　WebObjectsによる実装

5. 複数アクセスデータからの訓練事例作成

C4.5の後継バージョンであるC5を使ったアクセスデータからの情報アクセス技術の実験を試みている．SPIRALに関しては，同一URLであっても，アクセス時刻が異なれば，そのコンテンツが別途オブジェクトデータベースに格納される．したがって，同一URLではあるが，複数の時刻のページデータにより訓練事例を作成することができる．さらに，ページ内のリンクとの組み合わせにより，より複雑な訓練事例が作成できる．

上記の訓練事例の作成に関して，Prologにより実装したSPIRALシステム(KDDとしてはC4.5を使用)では，システムに組み込み済みの述語を用いた対応をのみであった．

WebObjectsを用いた今回の実装では，C4.5の後継バージョンであるC5をKDDツールとして使用することで高速性を確保し，大規模データに対応できるようにした．訓練事例の作成に関しては，ビジネスロジック部分で基本機能を与え，機能追加・拡張分に関しては，そのクラスを拡張することで対応でき，より柔軟なシステムとなっている．複数アクセスデータからの訓練事例作成に関して，バリエーションを豊富にすることができ，多様な実験をサポートすることが可能となった．

6. むすび

インターネットにおける知的情報統合利用環境SPIRALについて述べた．SPIRALでは，ユーザがアクセスしたWebサイトのURLだけではなく，そのコンテンツもタイムスタンプをつけてアクセスデータとしてデータベースに保存される．このデータベースに保存されるデータを活用することで，時間を考慮した知的情報アクセス技術の獲得に対応することができる．

Prologで実装していた既存システムにおけるパフォーマンスの問題により，大規模データを使った実験に対応できなかった点に関し，開発環境をWebObjectsとし，メインの開発言語をJavaとすることでその対応を図った．また，複数データからの訓練事例作成に関しても，基本機能が与えられたクラスの拡張により，より柔軟に行えるようになった．

今後，本実装方式により，大規模な実験を行っていく予定である．

謝辞

本研究は，日本学術振興会科学研究補助金（課題番号13780263）の補助を受けている．ここに謝意を示す．

文　献

[1] Seng Wai Loke, Andrew Davision: “LogicWeb : Enhancing the Web with logic programming”, Journal of Logic Programming, Vol. 36(3), pp.195-240,1998,.

[2] Apple Computer, Inc. : “Cyberdog 2.0（日本語版）ユーザーズガイド”, 1997.

[3] 松井英昭, 後藤文太朗: “WWWのシームレスな利用環境”, 人工知能学会研究会資料SIG-KBS-9803, pp.37-41, 1999.

[4] J.R.キンラン著, 古川康一監訳: “AIによるデータ解析”, トッパン, 1995.

[5] 石川雅弘，後藤文太朗："WWW アクセス活動と Web コンテンツの情報統合 "，情報処理学会第 60 回全国大会講演論文集，2000.

[6] 伊藤聖吾，後藤文太朗："Web の個人利用における編集活動とマウスアクションの拡張 "，情報処理学会第 60 回全国大会講演論文集，2000.

[7] 渥美尚紘，後藤文太朗："SPIRAL への C4.5 による学習モジュールの統合 "，情報処理学会第 60 回全国大会講演論文集，2000.

[8] 日立ソフトウェアエンジニアリング（株）："IntelligentPad 2.0" ,IntelligentPad 2.1bj 開発キッド付属 Document (1996)

[9] 石川雅弘，後藤文太朗："WWW アクセス活動と Web コンテンツの情報統合における履歴抽出精度の向上とその応用 "，情報処理学会第 62 回全国大会講演論文集，2001.

[10] 岡田哲弘，後藤文太朗：" リンク構造と時間軸を利用した Web ページ間の関連づけと学習の適用 "，情報処理学会第 62 回全国大会講演論文集，2001.

[11] Apple Computer, Inc. : "WebObjects 5 概要 "，2002.