情 報 学 基 礎 51－1
自 然 言 語 処 理 127－1
（1998. 9. 17）

WorkWare:WEBを用いた文書の時間順整理の試み

内野 寛治　津田 宏　松井 くにお

(株) 富士通研究所 ドキュメント処理研究部
{uchino,htsuda,kunio}@flab.fujitsu.co.jp

インターネットの普及によりネットワーク上には電子化された文書が爆発的に増加している。著者らは、ネットワーク上の文書群（電子メール、ニュース、WEB文書）をグループ内のユーザで共有するためのツールとして *WorkWare* を試作した。WorkWareでは、複数のビューをユーザに提供することで必要な情報を効率的に見つけられるような工夫をしている。本稿では、WorkWareの主な機能である文書群の時間順整理とその評価を中心に紹介する。

WorkWareの特徴の1つであるマルチビューでは、カレンダービューや超整理法ビューと呼ばれる異なった時間観点で文書群を整理するビューを複数用意し、ユーザは個々の目的に応じたビューを選択することで必要な情報を素早く探し出すことができる。もう1つの特徴である文書の取り込みインタフェイスについては既存のメーラなどのツールをそのまま利用できるよう工夫している。さらに取り込まれた文書は汎用的なXMLフォーマットでサーバ内に保存する。

システム評価では、日付情報の抽出精度やWorkWare内の文書の再利用率などを実際に取り込まれた文書やシステムの運用ログから分析してWorkWareで行なっている文書群の時間順整理の有効性を示す。

WorkWare:A WWW-based Chronological Document Organizer

Kanji Uchino　Hiroshi Tsuda　Kunio Matsui

Fujitsu Laboratories Ltd. Document Processing Laboratory
{uchino,htsuda,kunio}@flab.fujitsu.co.jp

To cope with constantly increasing use of digital documents, we developed *WorkWare*, a tool which integrates, shares, and organizes flow documents such as e-mail, news, HTMLs, and scheduled events. WorkWare provides multiple ways to access shared documents through WWW browsers.

This paper focuses on the views to organize documents chronologically. In one of the view called the calendar view, links between dates, schedules, and document contents are automatically extracted and shown in a calendar. This view offers an *episodic search* and *temporal push* interfaces to help users access desired documents. In another view called the stack-index view, documents are aligned with their latest accessed time for each user.

To create those views, natural language processing and data mining techniques are used. Users can directly capture documents into WorkWare from editors, mail/news readers, Windows desktops, and WWW browsers. Captured documents are shared as XML files in a server.

In the system evaluation, we calculate the rate of reuse and the reuse cycle of WorkWare documents by analyzing access-log and documents of WorkWare, and then show the efficiency of chronological organization.

1 はじめに

インターネットの爆発的な普及によって、ユーザは電子メール、ネットワークニュースやWWWページのようなフロー情報を扱う機会が増加している。とくにビジネス環境ではこれらの情報を効果的に収集/整理し仕事を進めることが必要だが、以下のような問題点があげられる。

- 日々扱うフロー情報の中にはゴミが多く、必要な情報だけを選択することが難しい。
- 必要な時に必要な情報を見つけられない（適当な情報の整理が行なえない）。
- 様々なメディアから情報は送られており、それぞれ異なったツールを使わなけらばならない。

著者らはこれらの問題点を解決するために、*WorkWare*と呼ばれる以下の特徴を持つシステムを試作した。

- グループ内のメンバによる情報共有。
- 文書から自動的に日付情報を抽出し文書群を複数の時間観点で整理するビュー。
- 日常使用しているメーラなどのツールをそのまま利用できるような文書取り込みインタフェイス。

最初の特徴に関しては、グループ内のメンバーで情報を共有することで、必要な情報のみをWorkWareに蓄積する。すなわち、メンバーという一種の「フィルタ」を通した情報のみを共有することで、ユーザのゴミ情報へのアクセスを防ぐ。2番目の特徴に関しては、多くのフロー情報（特にビジネス文書）を調査した結果、多くの文書中に日付表現が含まれていた。そこで、WorkWareでは自然言語処理やデータマイニングの技術を応用して日付情報を自動的に抽出し、文書を整理するための共通の軸として時間を利用している。具体的には、「絶対時間」、「相対時間」、「エピソード時間」、といった3つの異なった時間観点から文書整理を行ないそれぞれ異なったビューとしてユーザに提供している。3番目の特徴に関しては、既存のツールをそのまま使えることで文書共有を行なうために「新たなツールの操作を習得する」というユーザの負担を軽減する。また、ユーザにアクセスされて多くの共有情報を蓄積することができなければ、WorkWareの価値は半減してしまう。多くのユーザに抵抗なく使ってもらうにはシンプルな操作で文書の取り込みなどができなければならない。WorkWareではクリック、ドラッグ&ドロップなど単純な操作で文書を取り込むためのインタフェイスを備えている。

第2章ではWorkWareの概要を示す。第3章ではWorkWareで行なっている文書の時間順整理の特徴について説明する。第4章ではWorkWareの文書を取り込むためのインタフェイスの特徴について説明する。第5章ではシステムの評価とWorkWare内の文書群の特徴を分析し、第6章では結論を示す。

2 WorkWareの概要

WorkWareの概要を図1に示す。

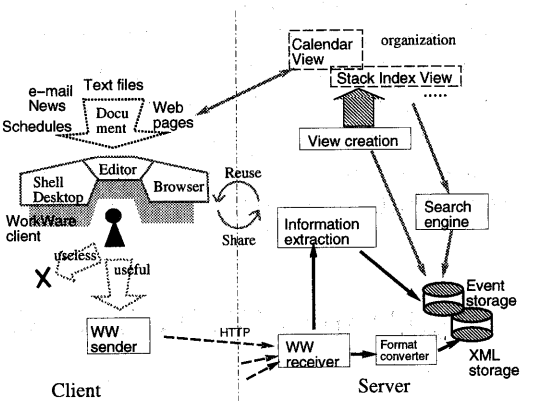

図1: WorkWareの概要

ユーザはメール/ニュースリーダやWWWブラウザなど様々なツールを用いてネットワーク上の文書にアクセスしている。WorkWareはこれらのツールをそのまま利用し、文書群を取り込み整理するメタツールとして位置付けられる。

ユーザは利用しているツール上の文書をクリックなどの簡単な操作で、システムに取り込むことができる。取り込まれた文書にはタイトルやジャンルなどのメタ情報が付加されHTTP (HyperText Transfer Protocol) を用いてWorkWare(WW)senderからWWreceiverへ送られる。サーバ内では取り込まれた文書からいくつかの2次情報を抽出し、XML[1]フォーマットに変換して保存する。

[1]http://www.w3c.com/XML/

2次情報の例としては、日付などの時間に関する情報や文書を特徴付けるためのキーワードなどがあげられる。

次にユーザが取り込んだ文書群にアクセスする場合、WorkWareではカレンダービュー、超整理法ビューなどのマルチビューを用意している。WorkWareに取り込まれた文書はこのような共有/再利用というプロセスをたどる。

3 文書群の時間順整理

1章でも説明したように、我々が日常扱うフロー情報の多くは日付情報を含んでいる。著者らは文書群を整理するための軸としてこの時間に着目し、以下の3つの観点で分類した。

絶対時間 文書内容に含まれている日時やその文書が作成/WorkWare内に取り込まれた日時を表す。

相対時間 ユーザがいつ文書にアクセスしたのかを表す。同じ文書であってもユーザによってこの値は異なる。

エピソード時間 文書中に含まれる日付に関連するスケジュールイベントを表す。ユーザの感覚の中では時間はスケジュールを単位に流れている場合が多い。

WorkWareではこれらの時間を軸にした文書整理をビューの形で実現している。

3.1 マルチビュー

マルチビューはWorkWareの文書アクセスに関する中心的な機能である。DB研究者の間ではビューやユーザビューという用語はオリジナルのデータとは異なった形式でユーザに表示する方法と位置付けられている。田中[4]はWWWへのビュー機能導入の必要性を示している。文書アクセスに関するWorkWareの基本的なアイデアは、共有文書群をユーザ側の様々な視点で参照できるようにするということである。図1で示したように、WorkWareのビュー作成モジュールはユーザのスケジュール情報とXMLのタグ情報を参照して様々なビューを作成する。ビューはCGIを用いてユーザが文書群を参照した場合に動的に作成される。

3.2 カレンダービュー

カレンダービューはWorkWareの基本的なビューであり、日付と文書とスケジュールイベントをリンク付けして表示する。すなわち、このビューでは絶対時間とエピソード時間で文書整理を行なう。また、ビューの最も重要な特徴であるスケジュールと文書の関連付けは4.1節で説明した日付情報の自動抽出プロセスを経て行なわれる。図2にカレンダービューの例を示す。

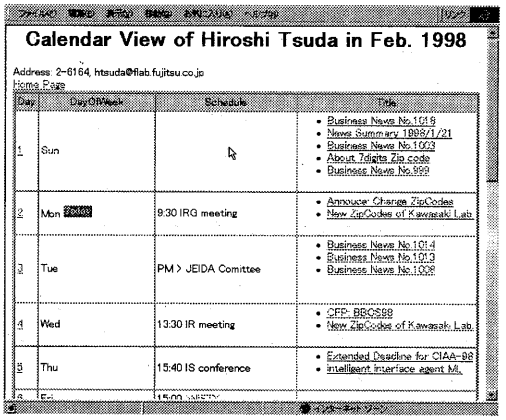

図2: カレンダービュー

HTMLのアンカー部をクリックすると、ビューを切替えるだけでなく、ユーザはグループ内の他のメンバーのホームページや共有文書の内容を参照することができる。すなわち、カレンダービューはイントラネット内のグループ共有情報へのポータルページ[2]として機能する。

図3はグループスケジュールのビューの例である。個人がいくつかのグループに所属する場合、グループやサブグループのスケジュールは自動的に継承され、継承されているスケジュールと個人のスケジュールは識別のため異なった色で表示される（図3の6月3日のスケジュールを参照）。また、個人やグループに予定がある場合はそれぞれの名前の頭文字がカレンダー上に表示される。

ビューの表示範囲は、週単位、月単位といった時間的な単位や個人やグループといった所属の階層を単位に切替えることができる。この機能よって、ユーザの

[2]ポートは港の意味。Yahooなどいろいろなサイトへの入口として機能するサイトはポータルサイトと呼ばれる

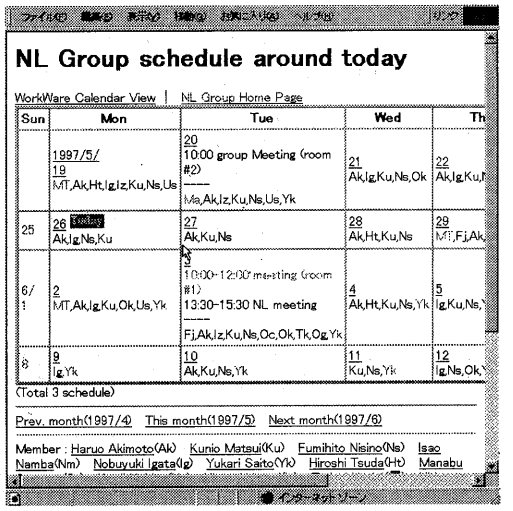

図 3: グループスケジュールのカレンダービュー

注目したいスコープでスケジュールを一覧することができる。カレンダービューにはエピソード検索、時制プッシュと呼ばれる 2 つの機能がある。

3.2.1 エピソード検索

カレンダービューでは全文検索のように文書に含まれるキーワードを検索キーにするのではなく、過去のスケジュールイベントに関連した文書情報検索が可能である。例えば、会議の資料や議事録など、会議というイベントに関連した資料を WorkWare でまとめていくとする。そして「確か、今年の 1 月の会議で書いた資料があったはずだが...」のように、数ヶ月前に作った資料を参照したい場合には、自分のスケジュールを遡りつつ、その日の前後の文書を探すことで、求める資料を得ることができる。個人的な経験や出会った人々に関する記憶 (認知心理学ではエピソード記憶と呼ばれる) の断片をもとに、エピソード時間で整理された文書群の検索が行えるわけである。

3.2.2 時制プッシュインタフェース

カレンダービューにおいて、文書は絶対時間で整理されるので、取り込んでから時間が経過して忘れてしまったような未来のイベントに関する情報でも、ユーザが個人のスケジュールをチェックすることで自然にそれらの情報を思い出すことができる。会議の案内や投稿論文の締切の案内などがそのいい例である。カレンダービューのこのような機能を時制プッシュと呼ぶ。従来のプッシュ型情報提示は、PointCast や Castanet に見られるように、ニュースのような速報性の高い情報を、興味ある人に提示するというケースが多かった。しかし、WorkWare の時制プッシュでは、以前知らせた情報をそれが必要な時に再度提示するという特徴を持つ。図 2の例では、「Extended Deadline for CIAA-98」のような論文の〆切日についての記事がカレンダー上に見受けられる。

3.3 超整理法ビュー

超整理法ビューは、相対時間で文書整理を行なう。具体的には、文書のタイトルリストを最近にアクセスされた順に並べ替えてユーザに提示するビューである。野口 [3] は文書整理について、文書を複数のクラスに分類することは出来ない、文書をどのクラスに分類したのかを忘れてしまう、などといった問題点を挙げている。その解決法として、文書を分類するのではなく、単にアクセスした順に並べ替えて文書整理する手法を提案している。この手法は以下のステップで実現される。

1. それぞれの書類を同じ大きさの封筒に入れ、その封筒を本棚に本を並べるように並べる。
2. 読んだばかりの書類が入った封筒を並んでいる封筒の一番左端に置く。
3. 書類を探す場合は左端の封筒から探し始める。

この手法では、新しい書類もしくは最も頻繁にアクセスされたものが常に並びの左端に置かれる。図 4は以上のステップを計算機上で実現した超整理法ビューの例である。ユーザが取り込んだ文書のタイトルがリストとして並べられている。ユーザはタイトルをクリックすることでその文書内容を読むことができ、その後このビューに再びアクセスすると以前読んだ文書のタイトルがリストのトップに表示される。ユーザがビューにアクセスすることにより「新しい」もしくは「頻繁にアクセスされる」文書のタイトルが自動的に常にリストの上位を占めるようになる。言い替えればユーザアダプテーションが自動的にこのビューでは実現されている。超整理法ビューはグループ内のどのユーザからもアクセスできるが、オーナー以外からアクセスされた場合はリストの順位の変更は行なわれない。

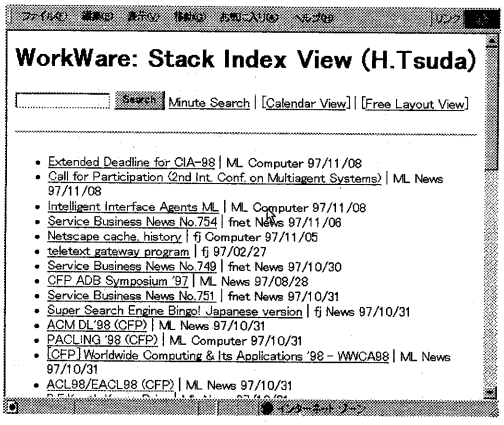

図 4: 超整理法ビュー

表 1: 日付表現の例

Pattern	Explanation
23 March 1997	英語の基本パターン
23 March 97	西暦の 2 桁表現
Jan.28–Feb.10,'97	2 カ月に跨る表現
Dec.27-Jan.14	2 年に跨る表現
平成 8 年 6 月 25 日	年号を用いた基本パターン
1996/06/25	数字のみのパターン
H8.6.25	年号が省略形
6 月中旬	慣用表現（15 日とみなす）
6 月下旬	慣用表現（月の最終日）

4 WorkWare における文書群の取り込み

WorkWare では、文書を効果的にシステムに取り込むために UNIX や Windows 上の汎用的なエディタ、WWW ブラウザなど既存のツールやデスクトップ上で、ボタンのクリック、ドラッグ&ドロップなどの簡単な操作で文書を取り込むためのインタフェイスを備えている。

取り込まれた文書には、ジャンルや公開範囲などのメタ情報がユーザによって付加される。さらに情報抽出モジュールによって日付などの 2 次情報が抽出され、それぞれの値は XML タグの値として保存される。以下の節では、情報抽出と XML 変換について説明する。

4.1 日付情報の自動抽出

取り込まれた文書は WorkWare 内の情報抽出モジュールによって日付情報やキーワードなどの 2 次情報を表層上のパターンマッチを用いて抽出する。

日本語または英語で書かれた 300 以上の電子メール / ニュース、WWW 文書などのフロー情報を調べた結果、100 以上もの日付に関連する表現を見つけた。表 1はそれら日付表現の例である。これらから日付に関する拡張正規表現を作り、情報抽出モジュール内で抽出パターンとして利用している。抽出された日付情報は表 2で紹介する XML の **RelDate** タグの値として保存される。

4.2 文書の共通フォーマット

WorkWare において、ユーザによって付加されたメタ情報や情報抽出モジュールで自動的に抽出された 2 次情報は、XML タグの値として保存される。表 2は、WorkWare で利用している XML タグの一覧である。**Genre** や **Range** の値は文書取り込み時に WorkWare クライアントで付加されるメタ情報である。**RelDate** や **KeyWords** はタグの値はサーバ内の情報抽出モジュールによって自動的に抽出される 2 次情報である。

5 システム評価

本章では WorkWare の主な特徴である日付情報の自動抽出の精度と WorkWare の取り込まれた文書の性質について分析した結果を説明する。

5.1 日付情報の抽出精度

WorkWare に取り込んだ文書群（2442 通）について日付情報がどれくらい含まれているかを実際に調べた。その結果を図 5に示す。

調査の結果、1 つ以上の日付情報を含む文書の割合は全体の 81% であり、1 文書中に含まれる日付情報数の平均値は 3.69 であった。この結果は WorkWare で扱う文書群（主にビジネス文書）を時間軸で整理する有効性を示している。19% もの文書は内容に日付情報を含んでいない訳だがこれらの文書も WorkWare 内ではアクセスされた日時で整理される。また、日付情報を含む文書から 121 通を無作為に選び、WorkWare によって日付情報がどれくらい正確に抽

表 2: WorkWare 内の文書で用いている XML タグ

Tag	Explanation
CurrentTime	Captured time
SrcTime	Document creation time
DocType	Document type
Media	Document media
From	Creator
UserID	User ID of capturer
UserName	Capturer
Title	Title
Genre	Genre
Range	Range of information
FileHeader	File name in the server
SystemName	Client machine
RelDate	Extracted related dates
KeyWords	Extracted keywords

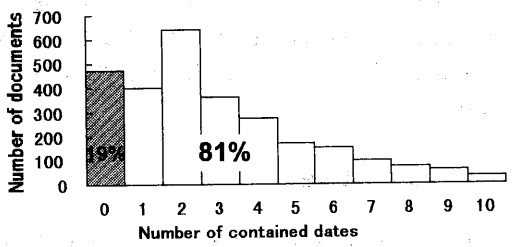

図 5: 日付情報を含む文書の割合

出されているかを調べたところ 94.4% の情報が正確に抽出されていることがわかった。

5.2 WorkWare 内の文書の再利用率

次に WorkWare 内に取り込んだ文書がどの程度グループのメンバーによって再利用されているかを WorkWare のアクセスログを解析して調べた。これによって WorkWare 内にグループでの共有価値の高い情報がどれくらい蓄えられているかを推定する。結果をグループのホームページと比較して表 3に示す。再利用率は WorkWare に取り込んだ文書とホームページからたどることができる文書のどの程度の文書数がユーザによってアクセスされたかを示す。再利用率は、被アクセス文書数 / 全文書数 で表される。

表 3: WorkWare とホームページの文書の再利用率

アプリケーション	再利用率
WorkWare	22.9% (560/2442)
ホームページ	28.5% (338/1184)

ホームページの方がわずかに再利用率が高いが、グループ内のメンバーに必要な情報を集めて人手と時間をかけてホームページを構築していることを考慮すると WorkWare で自動的に整理した文書群がこれに近い割合でアクセスされているという事実は、WorkWare による手軽な情報共有機能の有効性を示している。

5.3 WorkWare 内の文書の再利用期間

最後に WorkWare に取り込んだ文書の性質として、取り込まれてからグループのメンバーに再アクセスされるまでの期間を調べた。その結果を図 6に示す。

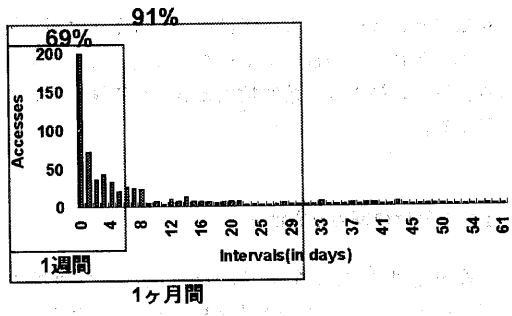

図 6: WorkWare 内の文書が再利用されるまでの期間

図より 2442 の文書の内 69% の文書が 1 週間以内に、91% の文書が 1 カ月間以内にグループメンバーによってアクセスされていることがわかる。すなわち、WorkWare で扱うフロー情報は本質的に再利用期間が短い情報であり、ホームページで扱うノウハウのような恒久的な情報ではないことを示している。

6 おわりに

本稿ではネットワーク上で増大する文書群の整理を行なうためのツールである WorkWare についての説明を行なった。本システムの主な特徴は文書共有、文書の時間順整理、文書取り込みのためのシンプルなインタフェイスである。

情報の時間順整理という点で WorkWare は Lifestreams [2] と似ているが以下の点で異なっている。

- WorkWare はグループのメンバー間での知識共有ツールであるが、Lifestreams は個人の文書整理ツールである。
- カレンダービューは文書から自動的に抽出した日付と文書間の関係を表している。このビューはストリームメタファを用いている Lifestreams では実現できない。多くのグループウエアでは日付に関するいくかのビューを備えているが、WorkWare の時制プッシュインタフェイスのような機能を持たせようとする場合、ユーザが明示的に文書をカレンダーの適当な位置に配置しなければならない。

また、WorkWare で扱う文書群の性質を調査した結果、情報共有のために運用しているグループのホームページに近い再利用率であることやほとんどの文書の再利用期間が 1 カ月以内という興味深い事実がわかった。再利用期間に関しては、例えば 1 カ月以上に渡ってアクセスされた文書などはホームページのノウハウ情報に移行し、1 カ月以内にアクセスされない文書は廃棄の対象とするなどの方法が改良点として考えられる。

最後に WorkWare ではディレクトリビューやフリーレイアウトビューと呼ばれるその他のビューも備えている。

ディレクトリビューは [7]、WorkWare 内の文書群に対して Yahoo[3]でサービスしているようなディレクトリを半自動で構築する。このビューはデータマイニング [1] 技術の応用であり、自動的に文書から抽出したキーワード間の共起関係を用いてディレクトリを構築している。ディレクトリ同士の関係はハイパーテキストのリンクを用いて表される。

フリーレイアウトビューは、WorkWare 内の文書タイトルを 2 次元空間上に整理するためのビューであり、ユーザによって自由に配置することができる。このビューはホワイトボードのメタファである。

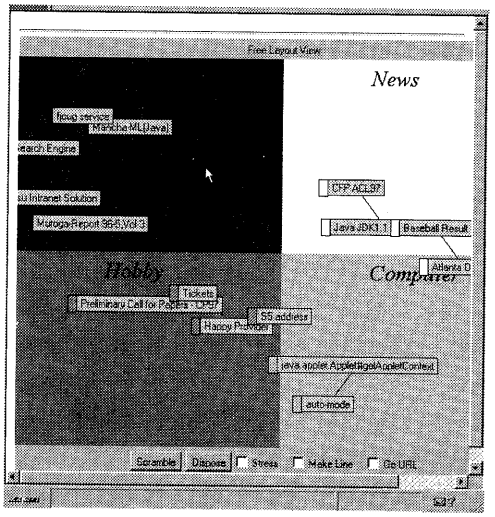

図 7: フリーレイアウトビュー

図 7はこのビューの例である。バックグランドの意味付けやタイトルノード間のリンクはユーザの目的によって自由に変更することができる。また、2 次元空間上での文書整理を容易にするため、リンクの張られていないタイトル同士は互いにに反発し合い、リンクの張られたタイトル同士は互い引合うように設計されている。

参考文献

[1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining*. MIT Press, 1996.

[2] E. Freeman and S. Fertig. Lifestreams: Organizing your Electronic Life. *AAAI Fall Symposium: AI Applications in Knowledge Navigation and Retrieval*, November 1995.

[3] 野口 悠紀雄「超」整理法. 中公新書, 1993.

[4] 田中 克巳 ネットワーク社会とマルチメディアデータベース. 情報処理学会学会誌, 38(1):24–29,1997.

[5] H. Tsuda, K. Uchino, and K. Matsui. WWW: WWW-based Chronological Document Organizer. *proc. of the 3rd Asia Pacific Computer Human Interaction*, 380–385, Jul. 1998.

[6] A. van Hoff, S. Shaio, and O. Starbuck. *HOOKED on JAVA*. Addison-Wesley, 1996.

[7] 津田 宏 テキストマイニングを用いた文書のディレクトリ整理. ソフトウエア科学会 *15* 回大会,1998.

[3] http://www.yahoo.com/