

卒業論文 2006 年度 (平成 18 年度)

デジタルスクラッピングシステム Markit の設計と実装

指導教員

慶應義塾大学環境情報学部

徳田 英幸

村井 純

楠本 博之

中村 修

高汐 一紀

湧川 隆次

慶應義塾大学 総合政策学部

川添 瑞木

zoe@sfc.wide.ad.jp

卒業論文要旨 2006年度(平成18年度)

デジタルスクラッピングシステム Markit の設計と実装

パーソナルコンピュータの普及、インターネットの発展により、デジタルメディアを用いたスクラッピングが注目されている。スクラッピングは、ある目的をもって情報を蓄積する行為である。従来のスクラッピングは、雑誌、新聞などの紙メディアを用いることが一般的であった。近年では、デジタルカメラ、スキャナの普及により、デジタルメディアを用いてスクラッピングできるようになった。このため、紙メディア、デジタルメディア双方の利点を生かしたスクラッピングシステムが一般的になると予想される。

スクラッピングは、1) 発見フェーズ、2) 蓄積フェーズ、3) 利用フェーズの3つの作業工程を経る。紙メディアとスキャナを利用したスクラッピングの場合、1) 新聞や雑誌を読む、2) 紙メディアをスキャナで読み込み、OCRによる文字情報を取得、メタデータの付加・分類、3) コンピュータで情報を検索、利用という工程を経る。このうち、紙メディアをスキャナで読み込み、OCRによる文字情報を取得、メタデータの付加・分類する作業は煩雑であり、利用者にとって大きな負担である。

本論文では、このような問題点を解決するデジタルスクラッピングシステムとして **Mark it** を提案する。Mark it では、紙メディアの記事に付箋を貼ることでデジタルメディアとして蓄積できる。これにより、利用者は興味を持った記事を容易にデジタルメディアとして蓄積できる。Mark it の評価としては、学習しやすさ、効率性、記憶しやすさ、間違えにくさ、主観的満足度の5つに関して、実験を行い評価した。

キーワード:

スクラッピング, 紙, デジタル, ユーザインタフェース, ユビキタス

慶應義塾大学 総合政策学部
川添 瑞木

Abstract of Bachelor's Thesis

Mark it : Research of Digital Sclapping

Recently, digital media are widespread by the improvement of the computer and the development of the Internet. Spread of digital media has made scrapping in digital possible, where in the past, it was only possible in paper media. Bookmarks on the web and scrapping files are one of the example. Both paper and digital have the strength and the weakness, and people choose digital or paper media for their need. Therefore various forms exist in handling both media in Scrapping.

While converting paper media to digital media in to the steps of scrapping, people tend to feel difficulties and the complexity to the operation. To give an example, in order to scrap by converting paper media into digital media, one needs to scan a paper media, then from the image information convert into to textual information by OCR, then sort by its category so it can be viewed easily, and add additional information to the scrap and storage it at last. It takes time overwhelmingly compared with the work in clipping scrapbook by paper media. However, the demand of newspapers and magazines rank high, and with the possibilities of digital media, which can duplicate massive information and search within easily, scrapping is very useful in the form of digital media.

In this research, the aim is to easily convert paper media to digital media, without any hard effort. In addition, objective will be building the system that shows human interests by marking. In this system, no hard effort are needed to scrap, during the steps of converting paper media into digital media.

Keywords:

Scrapping, paper, digital, user interface, ubiquitous

Mizuki Kawazoe
Faculty of Policy Management Keio University

目次

第1章 序論	1
1.1 はじめに	2
1.2 本研究の目的と意義	2
1.3 本論文の構成	3
第2章 今日のスクラッピング	4
2.1 情報の収集と蓄積	5
2.2 スクラッピングの定義	5
2.2.1 スクラッピングの工程	5
2.3 スクラッピングとメディア	7
2.3.1 紙メディアとデジタルメディア	8
2.3.2 スクラッピングにおける紙メディアとデジタルメディアの比較	8
2.4 スクラッピングにおけるメディアの使い分け	9
2.5 問題意識	11
2.6 本章のまとめ	12
第3章 印付け動作を用いた紙発見デジタル蓄積型スクラッピング手法	13
3.1 紙発見デジタル蓄積型スクラッピング	14
3.1.1 紙発見デジタル蓄積型スクラッピングの定義	14
3.1.2 蓄積フェーズにおける作業工程	15
3.1.3 利用フェーズにおける利用形態	16
3.1.4 機能要件	17
3.2 関連研究	18
3.2.1 やさしくデジタルファイリング	18
3.2.2 EnhancedDesk	18
3.3 印付け動作を用いた紙発見デジタル蓄積型スクラッピング手法	19
3.3.1 システムの動作概要	19
3.3.2 蓄積フェーズにおけるアプローチ	21
3.3.3 利用フェーズにおけるアプローチ	22
3.4 本章のまとめ	23
第4章 Mark it の設計	24
4.1 Mark it	25
4.1.1 Mark it の概要	25

4.1.2	想定環境	25
4.1.3	ハードウェア構成	25
4.1.4	ソフトウェア構成	26
4.2	各モジュールの設計	27
4.2.1	記事情報データベース	29
4.2.2	スクラッピング記事情報データベース	29
4.2.3	共有スクラッピング記事情報データベース	30
4.2.4	発見検知・蓄積モジュール	30
4.2.5	ビューワ管理モジュール	31
4.3	本章のまとめ	32
第5章	Markitの実装	33
5.1	実装環境	34
5.2	各モジュールの実装	34
5.2.1	データベース	34
5.2.2	発見検知・蓄積モジュール	35
5.2.3	ビューワ管理モジュール	36
5.3	本章のまとめ	37
第6章	評価	39
6.1	基本性能評価	40
6.1.1	検知所要時間	40
6.1.2	正答率	40
6.2	ユーザビリティ評価	41
6.2.1	実験	42
6.2.2	実験結果	44
6.2.3	考察	44
6.3	本章のまとめ	46
第7章	結論	47
7.1	本論文のまとめ	48
7.2	今後の課題と展望	48
7.2.1	付箋の検知における精度の向上	48
7.2.2	紙メディアからの情報取得	48

目次

2.1	様々なスクラッピング	5
2.2	スクラッピングの作業工程	6
2.3	メタ情報付加の例	7
3.1	紙メディアからデジタルメディアへのスクラッピングの作業工程	15
3.2	EnhancedDesk	19
3.3	システム利用の流れ	20
3.4	印付けの例	21
4.1	ハードウェア構成図	26
4.2	ソフトウェア構成図	28
4.3	発見検知・蓄積モジュール シーケンス図	30
5.1	ER 図	35
5.2	GAUDI	36
5.3	Baby Steps Attractor Loop	36
5.4	ARToolkit	36
5.5	ビューワ スクラッピング記事表示時	38
5.6	ビューワ 関連記事表示時	38
6.1	検知所要時間の測定結果	40
6.2	評価に用いた雑誌	41
6.3	実験環境	42
6.4	実験で用いたアンケート	43
6.5	作業時間平均	44
6.6	被験者が作業に要した時間の平均値, 最大値, 最小値	44

表 目 次

2.1	紙メディアとデジタルメディアの特徴	8
2.2	スクラッピングにおける紙メディアとデジタルメディアの長所と短所 . . .	10
4.1	Article テーブル	29
4.2	MarkedArticle テーブル	29
5.1	Markit サーバ及びユーザ PC 実装環境	34
5.2	ビューワ管理モジュール実装環境	34
6.1	貼付検知における正答率	41
6.2	アンケートの結果	45

第1章 序論

本章では，本研究の目的と意義，本論文の構成を述べる．

1.1 はじめに

近年、インターネットの普及とコンピュータの一般化により、人々はデジタルメディアを頻繁に利用するようになった。Webで情報検索を行うことや、Webサービスの利用、デジタルファイルの作成や共有は一般的な行為となっている。

そのような環境のなか、様々なメディアから興味深い情報を切り出して蓄積する行為であるスクラッピングが変化している。従来は、雑誌の切り抜きのように紙メディアを対象とした行為であったが、デジタルメディアの一般化により、デジタルメディアを対象としたスクラッピングが行われるようになった。例えば、Webのブックマークやファイルの保存が行われている。一方で、紙メディアも情報取得の方法として盛んに利用されており、紙メディアを対象としたスクラッピングも行われている。

紙メディアは閲覧性に優れており情報リテラシがなくても扱える一方で、デジタルメディアは検索性に優れており情報の更新も容易に実現できる。そのため、人々は利用目的や嗜好に応じて、スクラッピングで扱うメディアを使い分けている。Webで発見した情報をスクラッピングする手法として、ブラウザのブックマークに保存する手法、印刷してスクラップブックに貼り付ける手法の両方が用いられる。また、雑誌で発見した記事を、スキャナで取り込んでデジタルメディアとして保存するというスクラッピングも頻繁に行われる。このように、現在のスクラッピングは、紙メディアとデジタルメディアの両方を様々な形態で扱っている。

本研究では、スクラッピングのなかでも紙メディアをデジタルメディアとして保存する形態に着目し、人が興味を持った際にとる動作である印付けを用いたスクラッピングシステムを提案する。

1.2 本研究の目的と意義

本研究の目的は、紙メディアからデジタルメディアへのスクラッピング作業を面倒な作業を伴わずに行うことである。新聞記事のスクラップブックやWebサイトのブックマークなど、様々な形態のスクラッピングが行われているが、それらのスクラッピングのなかで、紙メディアをデジタルメディアへ変換して蓄積する形態のスクラッピングは、煩わしい作業を伴うため人々は気軽に行うことが出来ない。しかし、雑誌や新聞などの紙メディアからの情報取得は盛んに行われ、デジタルメディアは多量の情報蓄積、容易な検索の実現という特性を持っているため、紙メディアで得た情報をデジタルメディアで保存する欲求は増加している。紙メディアで発見した情報をデジタルメディアへ変換して蓄積する形態のスクラッピングを容易に実現可能にすることで、ユーザは自由な形態でスクラッピング作業を行うことが出来るようになる。

1.3 本論文の構成

本論文は、第2章においてスクラッピングについて考察し、問題意識を述べる。第3章では、本研究の対象とする紙メディアからデジタルメディアへのスクラッピングについて考察し、機能要件をまとめる。さらに、本研究のアプローチを説明する。第4章でシステムの設計、第5章で実装について述べる。第6章では本システムの評価を行い、第7章において本研究をまとめる。

第2章 今日のスクラッピング

本章では、ある目的をもって情報を蓄積する行為であるスクラッピングについて述べる。はじめに、情報の収集と蓄積を概観する。ついで、本論文におけるスクラッピングを定義し、スクラッピングの工程を整理する。その後、スクラッピングの対象となるメディアを整理する。つづけて、具体的なシナリオを挙げ、スクラッピングの問題点を導く。

2.1 情報の収集と蓄積

グーテンベルクによる活版印刷の発明以来、人々の周りには様々なメディアが登場し、情報が溢れている。そのような環境のなかで人々は、自らに必要な情報をメディアから収集し蓄積する行為であるスクラッピングを頻繁に行っている [1]。従来のスクラッピングは雑誌や新聞の切り抜きといった紙メディアに特化した行為であったが、近年のデジタルメディアの普及に伴い、web サイトのブックマークや PC 内へのファイルの保存といった新たな形態のスクラッピングが登場した (図 2.1)。

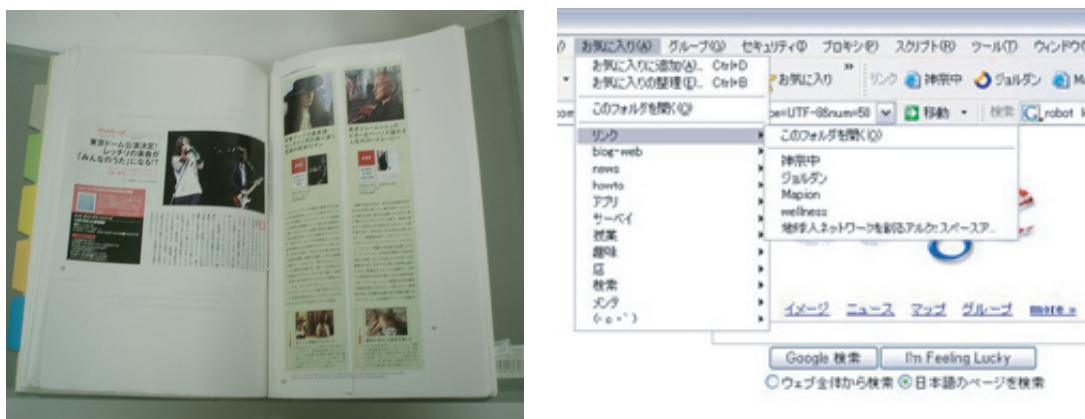


図 2.1: 様々なスクラッピング

2.2 スクラッピングの定義

スクラッピングとは、ある目的を持って情報を蓄積する行為である。例として、雑誌の切り抜きや web サイトのブックマークが挙げられる。

2.2.1 スクラッピングの工程

スクラッピングの作業工程は、人が様々な情報を取得するなかで興味深い情報を発見し、蓄積するという流れになる。本研究ではスクラッピングの工程を、発見と蓄積、利用の3つのフェーズにまとめる。図 2.2 に示す。

発見フェーズ

雑誌や新聞、web などの様々なメディアから情報を取得し、興味のある情報を発見するフェーズである。

発見フェーズにおける情報の発見には、以下の2種類がある。漠然と情報を取得するなかで興味深い情報を発見する場合と、目的を持って情報を検索し目当ての情報を発見する

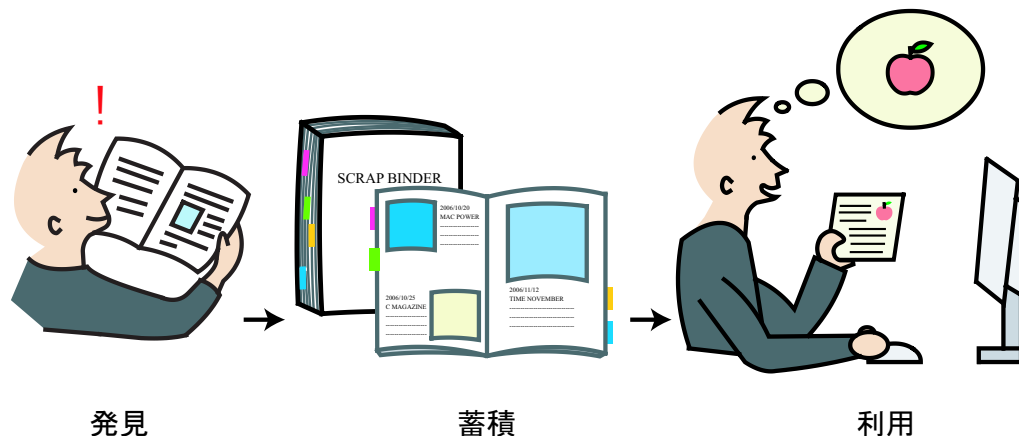


図 2.2: スクラッピングの作業工程

場合である。前者の例としては、新聞を読んでいると面白そうな本の書評記事を発見することが挙げられる。後者の例としては、購入予定のコンサートチケットの入手方法を web で検索し発見することが挙げられる。

蓄積フェーズ

発見フェーズで発見した情報を、蓄積するフェーズである。具体例として、記事の切抜きをスクラップブックに貼り付けることや、ブラウザのブックマークに URL を追加する行為が挙げられる。

蓄積フェーズにおいて、人はスクラッピングした情報を後に利用する際の利便性を考慮し、情報の分類とメタ情報の付加を行う。

- 情報の分類

ブラウザのブックマークをカテゴリ別に整理するように、蓄積する情報を分類する行為である。蓄積された情報が多量になると、後に利用する際に目的の情報を発見することが困難になるため、分類によって情報の検索性を高めている。

- メタ情報の付加

記事の掲載元やスクラッピングの際に注目したことなどの、スクラップ対象である記事のメタ情報を付加する行為である。例えば、紙のスクラップブックにおいて図 2.3 のように記事の掲載元やコメントなどの書き込みを行うことである。

スクラッピングでは、記事を掲載されていた場所から切り取って用いるため掲載元や記事の発行年月などの情報が失われてしまう。また、スクラッピングされた記事を利用する際には、その記事をスクラッピングした時から時間が経っているために、自らがスクラッピングした際の思考を忘れてしまっている。しかし、スクラッピングされた記事を利用する際には、記事の掲載元や発行年月、スクラッピングした際

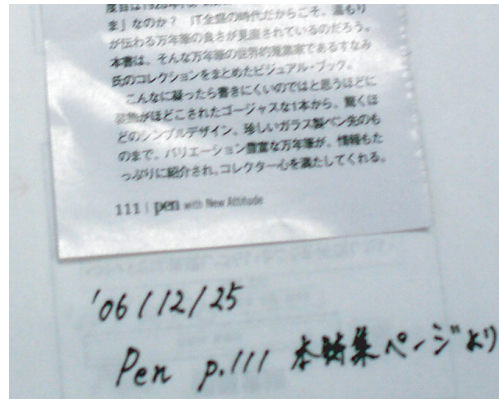


図 2.3: メタ情報付加の例

に注目していた点などの記事のメタ情報が重要であるため、人は各記事を蓄積する際にメタ情報を書き込むことで後の利用の際の利便性を確保している。

利用フェーズ

蓄積した情報を利用するフェーズである。例としては、関連情報の検索や情報の共有、蓄積された情報の見返しなどがある。

利用形態は、スクラッピングの目的によって変化する。好きな俳優に関する情報のスクラッピングのように、情報の収集自体が目的である場合は、利用形態は蓄積した情報の見返しとなる。また、気になるニュースのスクラッピングのように興味深い情報を発見した際のスクラッピングでは、スクラッピングした情報の関連情報を検索することが考えられる。

また、利用形態が関連情報の検索である場合は、検索によって発見した情報はまたスクラッピングの対象となり、次々とスクラッピング情報が増えていくこととなる。

スクラッピングは、後の利用のためになされるものなので、蓄積フェーズでは先に述べたように分類やメタ情報の付加など、利用フェーズでの利便性を考慮して蓄積が行われる。

2.3 スクラッピングとメディア

近年ではスクラッピングの対象となるメディアには紙メディアだけでなく、デジタルメディアが含まれている。メディアが異なる場合、スクラップブックへの貼り付けとブラウザのブックマークへの追加のように、スクラッピングの形態も異なる。本節では、紙メディアとデジタルメディアの特徴を比較し、メディアによるスクラッピングの特徴を述べる。

2.3.1 紙メディアとデジタルメディア

面谷信は、ディスプレイと紙の作業効率実験を行い、ユーザの主観評価の結果から人はディスプレイより紙の方が見やすいと感じると述べており、その原因として水平画面であることや周囲との明るさが均等であることなどを挙げている [2]。また Jakob Nielsen は、印刷デザインと web デザインの違いについて述べた文の中で、デジタルメディアはインタラクティブ性が優れていると述べている [3]。

このように、紙メディアとデジタルメディアにはそれぞれ特徴がある。表 2.1 に紙メディアとデジタルメディアの特徴をまとめる。

表 2.1: 紙メディアとデジタルメディアの特徴		
	紙メディア	デジタルメディア
長所	閲覧性が高い 情報の所有感が高い 書き込みが容易 電源が不要 記載場所の記憶が容易 軽く扱いやすい 情報リテラシーが不必要	検索が容易 情報の更新性 動画・音声を扱える 多量の情報が扱える 情報へのアクセスが多様 インタラクティブ性
短所	かさばる 劣化が激しい 更新性がない 動画・音声は扱えない	電源が必要 記載場所の記憶が困難 ディスプレイの枠内でしか閲覧できない 閲覧性が低い

2.3.2 スクラッピングにおける紙メディアとデジタルメディアの比較

両メディアの特徴をもとに、対象とするメディア別にスクラッピングの各フェーズの例と特徴を述べる。

- 紙メディアでの発見

紙メディアでの発見の例としては、雑誌や新聞で情報を取得するなかでの発見が挙げられる。第 2.3.1 項でまとめたように、紙は閲覧性が高いため、目的を持たずに情報を取得する場合には適したメディアである。しかし、検索性には優れていないため、目的を持って情報を検索する場合には適していない。

- デジタルメディアでの発見

web サイトや電子ブックにおける情報の発見が例として挙げられる。デジタルメディアは紙メディアとは反対に、閲覧性は低いが見つけやすさに優れている。そのため、目的

を持たずに長時間情報を取得する場合には適していないが、目的の情報を発見する場合は適している。また、動画や音声の情報を発見することが可能である。

- 紙メディアでの蓄積

紙メディアでの蓄積の例としては、雑誌や新聞の切り抜きをスクラップブックに保存することが挙げられる。紙メディアは検索性や復元性に乏しいため、情報が大量になった際に整理することが困難である。また、劣化の問題と大量の情報を扱えないという問題があるため、情報の蓄積には基本的に向かない。しかし、デジタルメディアよりも紙メディアのほうが所有感があるため、母親が息子の活躍をスクラップリングする場合のように、情報の収集自体が目的となるスクラップリングでは、紙メディアでの蓄積は適している。

- デジタルメディアでの蓄積

webサイトのURLをブックマークすることや、情報をファイルに保存することが例として挙げられる。デジタルメディアは多量の情報を扱うことが出来、検索も容易であるため、多量の情報を整理して蓄積するのに適している。情報の更新も容易である。また、利用フェーズにおいて、デジタルメディアとしても紙メディアとしても容易に利用できる。しかし、記事のように情報の意味が重要な紙メディアではなく、写真や絵など閲覧が目的となる紙メディアをスクラップリングする際には、閲覧性の低いデジタルメディアは適していない。

- 紙メディアでの利用

紙メディアでの利用の例としては、スクラップブックを見返すことや、雑誌のコピーやwebサイトの印刷を友人と共有することが挙げられる。紙メディアは閲覧性に優れているため、スクラップした情報を見返すことには向いている。しかし、検索性は低いため利用の際に目的の情報を発見することが困難である。

- デジタルメディアでの利用

検索エンジンでの関連情報の検索や、友人とのファイルの共有が例として挙げられる。関連情報の検索や友人との情報共有には、検索性が高く複製が容易なデジタルメディアは適している。しかし、目的はなく漠然とスクラップした情報を見返したい場合や、じっくりと文章を読みたい場合には、閲覧性が低いデジタルメディアは適していない。

表 2.3.2 に、スクラップリングにおける紙メディアとデジタルメディアの長所と短所をまとめる。

2.4 スクラップリングにおけるメディアの使い分け

前節で述べたように、スクラップリングは扱うメディアによって特徴が異なるため、人々は両メディアを目的や嗜好によって使い分けている。例えば、発見フェーズと蓄積フェーズ

表 2.2: スクラッピングにおける紙メディアとデジタルメディアの長所と短所

		長所	短所
発見	紙	目的のない閲覧で軽負担	目的のある閲覧で検索が困難
	デジタル	目的のある閲覧で検索が容易 動画や音声扱える	長時間の閲覧で負担
蓄積	紙	所有感がある	整理が困難 劣化が激しい
	デジタル	整理が容易 多量に保存できる 更新が容易	閲覧目的の蓄積に向かない
利用	紙	目的のない閲覧で軽負担	検索が困難
	デジタル	目的のある見返しに向く 関連情報の検索が容易 複製が容易	目的のない見返しで負担

では紙メディアを用いて利用フェーズではデジタルメディアを用いる場合や、発見フェーズではデジタルメディアを用いて、蓄積フェーズは紙メディアで行い、利用フェーズは再びデジタルメディアで行う場合など、様々な場合が存在する。以下に、様々な形態のスクラッピングのシナリオを挙げる。

シナリオ 1：紙発見・紙蓄積・紙利用

太郎さんは、ガーデニングを趣味としている。ある日いつものようにガーデニング雑誌を読んでいると、新種のひまわりに関する記事が載っていた。興味を持った太郎さんは記事を切り抜き、ガーデニングに関するスクラップブックに貼り付け、雑誌名と号数を記録した。また、ガーデニング仲間の次郎さんにひまわりについて話すと興味を持ったため、記事のコピーを手渡した。

シナリオ 2：デジタル発見・紙蓄積・デジタル利用

花子さんは現在、大学の卒業論文を書いている。webで関連研究を検索していると、自分と同じ問題意識を持った論文を発見した。関心を持った花子さんはその論文を印刷し、関連研究のファイルに追加した。更に関連研究を調べるために、その論文の参考文献に挙げられている論文をwebで検索した。

シナリオ 3：紙発見・デジタル保存・デジタル利用

みちこさんは映画鑑賞が趣味であり、映画関連のコラムはデジタルメディアとして大量

にスクラッピングしている。映画雑誌を読む際には、気に入ったコラム全てに付箋を貼って印をつけておき、後日貯めた記事をまとめてデジタルメディアに変換している。変換の手順は、該当記事をスキャンし、OCRを利用した画像ファイルからPDFファイルへの変換ソフトを用いてPDFファイルへの変換を行い、記事が掲載されていた雑誌名と号数をPDFファイルに書き込み、分類して保存するという流れになっている。みちこさんはこの手順が面倒であり、雑誌を読む作業を中断するのが嫌なため、まとめてスクラッピングすることになっている。

2.5 問題意識

本節では、紙メディアとデジタルメディアの両方を用いるスクラッピングにおける問題点を挙げる。第2.4節のシナリオを参考に、メディア別のスクラッピング形態における問題点について考察する。

一般的に、各フェーズを同一メディアで行う形態よりも、フェーズ間で扱うメディアが変化する場合の方が、ユーザは作業に負担を感じる。メディアには各々対応する作業形態があるために、扱うメディアが変化する際には作業形態もそれに伴い変更しなければならない。そのため、作業工程に隔たりが出来てしまいユーザは面倒を感じてしまう。メディアが変化する形態としては、デジタルから紙への変化と紙からデジタルへの2種類がある。以下に特徴を述べる。

- デジタルから紙への変化

デジタルメディアから紙メディアへ変化する場合の問題点としては、デジタルメディアは動画や音声のように扱える情報の種類が多いため、紙メディアには変換できない情報があることが挙げられる。しかし、第2.4節で述べたように、人々は目的や状況によってメディアを使い分けており、動画や音声を利用したい場合は紙メディアでなくデジタルメディアを利用すると考えられるので、動画や音声が扱えないことは重大な問題とはならない。また、文字と画像の静的な視覚情報をデジタルメディアから紙メディアへ変換する手法としてはプリンタが定着しており、人々は手軽に変換を行うことが出来ている。

- 紙からデジタルへの変化

紙メディアからデジタルメディアへの変換を行うためには、現在ではシナリオ3でのみちこさんのように、手で書籍を持ちページをめくるといった紙メディアを扱う作業を中断し、PCでデジタルメディアを扱う際のマウスとキーボードによる作業を行う必要がある。このように作業の形態が変化する場合、スクラッピング作業におけるユーザの負荷は大きくなる。

さらに、紙メディアからデジタルメディアへの変化が、発見フェーズと蓄積フェーズ間で生じる場合は、スクラッピング対象の情報を紙からデジタルへ変換する作業だけでなく、記事の分類とメタ情報の付加という作業も行わなければならない。その

ため、この段階で作業工程に隔たりが生じてしまうと、ユーザへの負荷はさらに大きくなる。

前述のように、現在のスクラッピング形態のなかで最も解決すべき問題点が多いのは、**発見フェーズを紙メディアで行い蓄積フェーズをデジタルメディアで行う形態**である。以降、この形態を**紙発見デジタル蓄積型**と呼ぶ。紙発見デジタル蓄積型スクラッピングは、現状では非常に面倒な作業であるため、コンピュータを用いた作業が苦手な人々や大量にスクラッピングを行う人々は、この形態でのスクラッピングを敬遠しがちである。しかし、表 2.3.2 に見られるように、発見フェーズでの紙メディアの利用には閲覧性の高さ、蓄積フェーズでのデジタルメディアの利用には、整理の容易さや情報の更新性など多くの利点がある。紙発見デジタル蓄積型スクラッピングを容易に行うことが出来れば、両者の利点を兼ね備えたスクラッピングを行うことが出来るようになる。よって本研究では、紙発見デジタル蓄積型スクラッピングを対象とする。

2.6 本章のまとめ

本章では、スクラッピングの定義と分類を行った。また、紙とデジタルの2種類のメディアの特徴をまとめ、スクラッピングにおけるメディアの使い分けについて述べた。さらに、本研究の問題意識を紙発見デジタル蓄積型スクラッピングとし、本スクラッピング行為では、作業形態の隔たりにより人に多大な負担を負わせることが問題であると述べた。

第3章 印付け動作を用いた紙発見デジタル蓄積型スクラッピング手法

本章では，本研究のアプローチである印付け動作を用いた紙発見デジタル蓄積型スクラッピングについて述べる．まず紙発見デジタル蓄積型スクラッピングに必要な手順について説明し，次に本研究で目指すべき手法に求められる機能要件を挙げる．次に，関連研究について述べる．その後，本研究のアプローチを述べ，最後に本システムの利用の流れを述べる．

3.1 紙発見デジタル蓄積型スクラッピング

本節では、紙発見デジタル蓄積型スクラッピングについて述べる。はじめに、紙発見デジタル蓄積型スクラッピングを定義する。次に、紙発見デジタル蓄積型スクラッピングの蓄積フェーズにおける作業工程を考察し、利用フェーズにおける利用形態を挙げる。

3.1.1 紙発見デジタル蓄積型スクラッピングの定義

紙発見デジタル蓄積型スクラッピングは、発見フェーズを紙メディアで行い、蓄積フェーズをデジタルメディアで行う形態のスクラッピングである。本項では、紙発見デジタル蓄積スクラッピングにおける発見、蓄積、利用の3つのフェーズの定義を行うことで、紙発見デジタル蓄積スクラッピングを定義する。

発見フェーズ

紙発見デジタルスクラッピングにおける発見フェーズでは、ユーザは雑誌や新聞などの紙メディアから情報を取得し、興味のある情報を発見する。発見フェーズにおいてユーザがすべきことは、一般的なスクラッピングと同様に、様々な紙メディアを読み、情報を取得することのみである。

蓄積フェーズ

紙発見デジタル蓄積における蓄積フェーズでは、ユーザは発見した紙メディアをデジタルメディアへ変換し、蓄積する。一般的なスクラッピングと同様に、ユーザは後に利用する際に検索のキーや記憶の補助として役立つように、デジタルメディアへと変換した情報に対してメタ情報を付加し、分類する。

メタ情報には、スクラッピング対象である情報自体の属性情報（雑誌名、発行年、ジャンル等）である静的なものと、ユーザが独自に付加して更新され続ける情報（お気に入り度、自分がなぜスクラッピングしたか、関連する情報等）である動的なものの2種類がある。

利用フェーズ

紙発見デジタル蓄積における利用フェーズでは、蓄積したデジタルメディアを印刷する場合のように紙メディアで利用されることもあるが、一般的にはデジタルメディアのまま利用される。

利用フェーズにおいて多く行われる行動は、見返しと関連情報の検索である。見返しとは、蓄積した情報を見返す行為である。関連情報の検索とは、蓄積した情報の関連情報を検索する行為である。ソーシャルブックマークのように、他のユーザがスクラッピングし

た情報内から関連情報を得る場合もある。また、得られた関連情報は、関連情報を検索する元となったスクラッピング情報のメタ情報と位置づけられる。

3.1.2 蓄積フェーズにおける作業工程

紙発見デジタル蓄積型スクラッピングにおいて、蓄積フェーズは、スクラッピング箇所の提示、メタ情報の付加、分類の3つの作業工程からなる。図3.1に、蓄積フェーズにおけるユーザの各作業工程と、各々に対応するコンピュータの処理を示す。

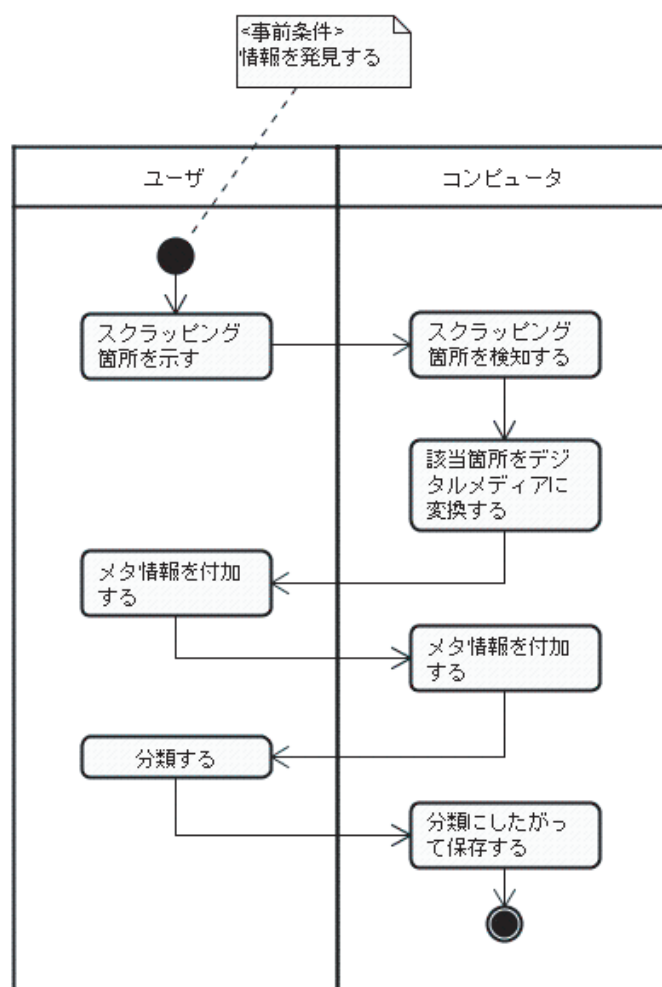


図 3.1: 紙メディアからデジタルメディアへのスクラッピングの作業工程

以下に、各工程について説明する。

- スクラッピング箇所の提示

まずユーザは、スクラッピングを行いたい箇所を何らかの方法でコンピュータに提示する。コンピュータは該当箇所を認識し、紙メディアからデジタルメディアへ変換する。

- メタ情報の付与

次にユーザは、デジタルメディアへ変換された情報に付加すべきメタ情報を入力する。デジタルメディアにはメタ情報が付与されていないので、後の利用時に不便を感じないようにするために行う。コンピュータは、ユーザが入力したメタ情報を、変換されたデジタルメディアに付与する。

- 分類

最後にユーザは、利用フェーズにおける利便性のために情報の分類を行う。コンピュータは、ユーザから示された分類に従って情報を保存する。

以上のように、ユーザは、スクラッピング箇所の提示、メタ情報の付与、分類という3つ作業によってデジタルメディアへ変換する。コンピュータは、ユーザの作業に対応して各情報の取得・保存を行う。この3点は、紙メディアからデジタルメディアへの変換を行う作業であるため、第2章において述べた作業工程の隔たりが発生しないように考慮する必要がある。

3.1.3 利用フェーズにおける利用形態

次に、紙発見デジタル蓄積型スクラッピングの利用フェーズにおける利用形態について述べる。利用フェーズにおいて多く行われる行為である見返しと関連情報の検索について、それぞれの考慮すべき点を以下に述べる。

- 見返し

見返しは、スクラッピングにおける最も基本的な利用形態である。考慮すべき点としては、見返した際の補助として、蓄積フェーズで付与したメタ情報を提示することが挙げられる。

- 関連情報の検索

関連情報の検索によって発見された関連情報は、検索の元となった情報と同じようにスクラッピングされる場合が多い。その際、関連情報は元の情報のメタ情報となり、関連情報の分類やメタ情報は元の情報に関連するものとなる。このような関連情報と元の情報との関連については、考慮する必要がある。

3.1.4 機能要件

上述の蓄積フェーズと利用フェーズに対する考察から、紙発見デジタル蓄積型スクラッピングを支援するシステムに求められる機能要件を整理する。まず蓄積フェーズにおける機能要件を述べ、次に利用フェーズにおける機能要件を述べる。

蓄積フェーズにおける機能要件

蓄積フェーズにおける機能要件は、以下の3点が挙げられる。

1. スクラッピング箇所の容易な伝達

紙メディアの情報をデジタルメディアへ変換するためには、ユーザは変換の対象となる箇所をコンピュータに対して容易に伝達できる必要がある。例えば、スクラッピング箇所を伝達するために、対象箇所を切り取ってスキャナで読み込むような作業が必要であれば、ユーザに大きな作業負担を与えてしまう。

2. メタ情報の容易な伝達

情報をスクラッピングする際には、スクラッピング対象の情報と共にメタ情報を保存する。そのため、保存すべきメタ情報を容易にコンピュータに伝達できる必要がある。また、掲載元やジャンルなど、記事自体の情報である静的なメタ情報は、コンピュータが自動取得できる必要がある。

3. 分類方法の容易な伝達

蓄積フェーズでは、スクラッピング情報の分類を行う。分類方法はユーザの目的や嗜好によって異なっており、自動で分類を行うことは出来ないため、分類情報はユーザから取得しなければならない。その際に、ユーザはできるだけ容易に分類情報を伝達できる必要がある。

これらの作業は、蓄積フェーズの作業工程に中断を伴わないように行わなければならない。

利用フェーズにおける機能要件

利用フェーズにおける機能要件は、以下の3点である。

1. 見返し行為の支援

ユーザが見返しを容易に行えるように、関連する情報はまとめて表示するなど、スクラッピングされた情報の表現方法を工夫する必要がある。特に、雑誌名や発行年などのスクラッピング対象である情報自体のメタ情報は、スクラッピング対象の情報と共に表示する必要がある。また、見返したいスクラッピング情報を容易に見つけられるように、ユーザが示した分類方法に基づいてきちんと分類されている必要がある。

2. 関連情報の検索・表示支援

ユーザが蓄積した情報の関連情報にすぐにアクセスできるよう、関連情報を容易に検索できる必要がある。また表示の際にも、元のスクラッピング情報との関連が、直感的にわかるように表示方法を工夫する必要がある。

3. 共有の支援

あるユーザが蓄積した情報を、他のユーザと簡単に共有できるようにする必要がある。また、共有情報のメタ情報に対する表示方法も、見返し行為の支援で述べたように、工夫する必要がある。

3.2 関連研究

本節では、既存の紙発見デジタル蓄積型スクラッピングシステムと本研究の関連研究を挙げる。

3.2.1 やさしくデジタルファイリング

やさしく新聞スクラップは、メディアドライブ株式会社が開発したデジタル蓄積型スクラッピングシステムである [4]。様々な形式のファイルを一括で管理することが出来、紙発見デジタル蓄積型スクラッピングにも対応している。スキャナで読み取った紙メディアは、OCR によって透明テキスト付き PDF へ変換することができる。整理や後の利用における利便性を重視しており、様々な機能が実装されている。例えば、ファイルの分類、日付やファイル形式によるツリー構造での表示、重要な箇所にマーカーを引いたりコメントの追加などができるアノテーション機能などである。

しかし、紙メディアからデジタルメディアへの変換を、OCR とスキャナを用いて行う手法では、蓄積フェーズでのにおいてユーザの負担が大きい。スキャナで読み込む作業だけでなく、分類とメタ情報付加もコンピュータを用いて行わなければならないため、紙メディアを読んでいる作業を大幅に中断するという問題点は解決されていない。

3.2.2 EnhancedDesk

小池らの提案する EnhancedDeck は、紙とデジタル情報の統合を目的とした机方実世界志向システムである [5]。机上の本や書類といった紙メディアに関連するデジタル情報を机上に投影することで、従来はユーザが行っている紙とデジタル情報の連携を自動的に行う。また、指によるデジタル情報の操作が可能であり、デジタル情報を直感的な動作で扱うことができる。図 3.2 にシステム使用図を示す。紙メディアの認識には 2 次元バーコードを、手指の認識には赤外線カメラ、デジタル情報の投影にはプロジェクタを用いる。

EnhancedDesk は、指を用いたデジタル情報の操作によって、現実世界のものである紙とデジタル情報の統合という目的を達成している。本研究は、紙発見デジタル蓄積型ス

クラッピングを対象としており，スクラッピングには，記事の分類やメタ情報の付加，スクラッピング記事の見返しといった，スクラッピング特有の作業がある．スクラッピングを支援するためには，それらのスクラッピング特有の作業に適したアプローチが必要となる．

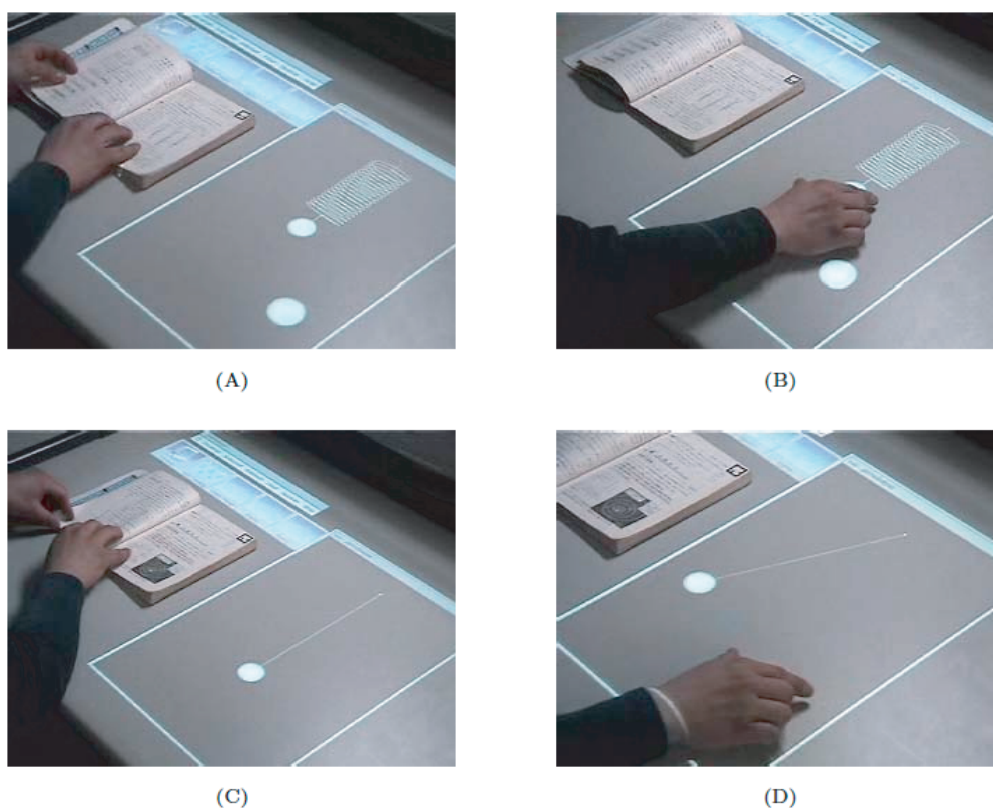


図 3.2: EnhancedDesk

3.3 印付け動作を用いた紙発見デジタル蓄積型スクラッピング手法

本節では，本研究で提案するアプローチである，印付け動作を用いた紙発見デジタル蓄積型スクラッピング手法について説明する．まず本手法の概要について説明し，その後で手法の詳細について説明する．

3.3.1 システムの動作概要

本研究では，紙発見デジタル蓄積型のスクラッピングを支援するための手法として，印付け動作を用いたスクラッピングを提案する．図 3.3 に，印付け動作を用いたスクラッピ

ング手法の利用の流れを示す。

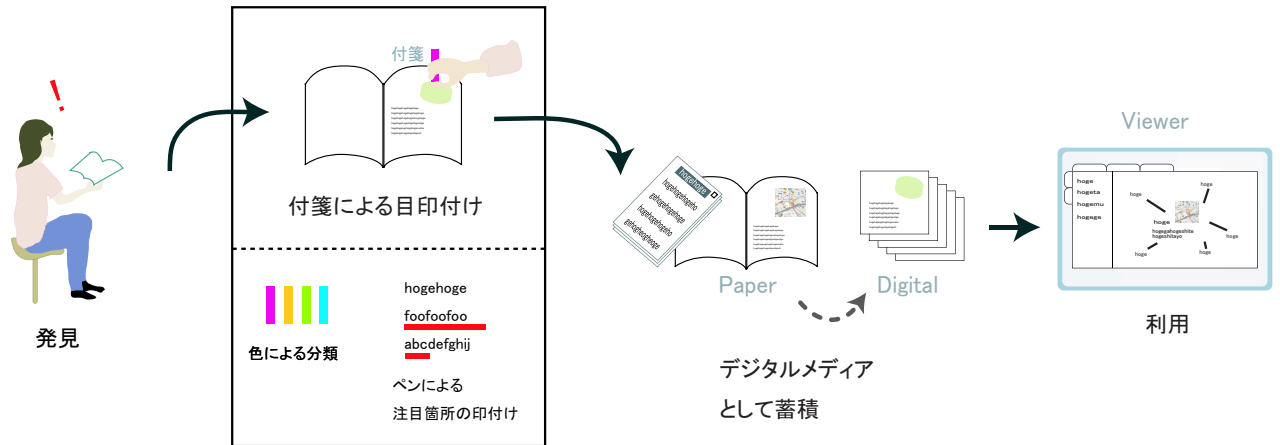


図 3.3: システム利用の流れ

印付け動作を用いたスクラッピング手法においてユーザーが行う作業について、スクラッピングの発見フェーズ、蓄積フェーズ、利用フェーズに分けて以下に説明する。

- 発見フェーズ

本手法における発見フェーズは、既存の紙発見デジタル蓄積型スクラッピングと同じである。ユーザは雑誌や新聞などの紙メディアで情報を取得する際に、気になる情報を発見する。

- 蓄積フェーズ

本手法における蓄積フェーズでは、ユーザは情報を蓄積する際の分類に対応した色の付箋を貼り、注目箇所を記しておきたい場合は注目箇所にペンで印をつける。この印付け動作によって、紙メディアはデジタルメディアへと変換され、蓄積される。

- 利用フェーズ

スクラッピングした情報は、ビューアを通して見返すことができる。このビューアは、スクラッピングした情報のメタ情報や関連情報も同時に表示する。また、スクラッピングした情報を他ユーザと共有する機能を有する。

以上のように、本研究で提案する手法では、ユーザがスクラップしたい記事をはさみで切り取る作業や、スキャナにかけた後にキーボードやマウスを操作してメタ情報を入力す

る作業は必要がない。また、スクラッピングされた情報を利用することに特化したビューアが存在するため、デジタルメディアの特性を活かしてスクラップ情報を有効活用できる。

次項より、第3.1節で述べたそれぞれの機能要件に対して、本研究が提案するアプローチの特徴を詳細に述べていく。

3.3.2 蓄積フェーズにおけるアプローチ

蓄積フェーズにおける、機能要件に対するアプローチの詳細について述べる。蓄積フェーズにおける機能要件は、スクラッピング箇所の容易な伝達、メタ情報の容易な伝達、分類方法の容易な伝達の3点であった。本研究で提案する手法では、上記の3つの機能要件に対するアプローチとして、付箋をトリガとしたスクラッピング箇所の伝達、ペンによる印付けを用いたメタ情報付加、付箋の色を用いた情報の分類、の3つを提案する。以下に、それぞれについて説明する。

付箋をトリガとしたスクラッピング箇所の伝達

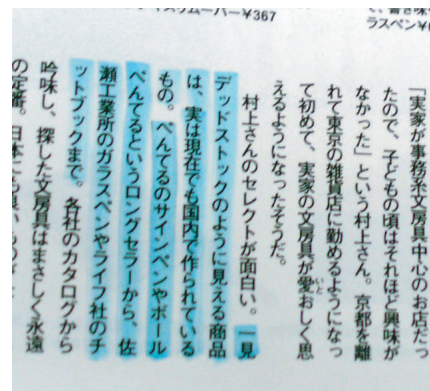
スクラッピング箇所の容易な伝達方法として、本研究では「付箋をスクラッピング対象箇所に貼り付ける」という動作を用いる。人々は、紙メディアのなかに注目した情報を発見した場合に、付箋によって該当記事への印付けを行うことで、その情報を特徴づけるという行為を頻繁に行う。よって、この動作を紙発見デジタル蓄積型の蓄積方法として利用することは、ユーザにとって直感的かつ容易であると考えられる。



付箋



折り目付け



ペンによる書き込み

図 3.4: 印付けの例

紙メディアに対して我々がよく行う印付け動作には、付箋の貼付だけでなく、葉を挟む動作やページに折り目をつける動作も挙げられる(図3.4)。しかし、スクラッピングの対象はページ単位ではなく、記事単位の場合もあるため、葉を挟む動作やページに折り目をつける動作では全てのスクラッピング対象箇所に対応できない。よって本研究では、付箋を貼るという行為をトリガとして用いた。更に後で詳しく述べるが、付箋は情報の分類とも親和性が高いという特徴も持つため、有用である。

ペンによる印付けを用いたメタ情報付加

メタ情報の容易な伝達方法として、本研究では「ペンによる印付け」という動作を用いる。ユーザは紙メディア上の注目した情報に対して、特に注目した点にペンで印付けすることが多い。本研究では、ペンで注目箇所を囲う動作をメタ情報の付加に利用することで、ユーザの自然な作業を妨げることなくデジタルメディアへのスクラッピングを行うことを実現する。

付箋の色を用いた情報の分類

分類方法の容易な伝達を実現するために、本研究では「付箋の色を用いた情報の分類」を行う。付箋をスクラッピング箇所の伝達に使うことは先に述べたが、この付箋の色を利用することで、マウスやキーボードを利用することなく分類方法をデジタルメディアに伝達することができる。また、付箋を貼るという1つの行為でスクラップ箇所の伝達と情報の分類方法の伝達の2つの作業を行えるため、ユーザの負担も軽減されることとなる。

情報の分類に付箋の色を用いる手法の問題点としては、色数の制限による分類項目の制限と、色と項目の対応を覚えなければならないという2点が挙げられる。色数の制限による分類項目の制限については、人は情報を階層化して分類することが多く、一番上の階層では分類数は多くない。一番大まかな分類を指定するのであれば、色数の制限は大きな問題ではない。色と項目の対応に関しては、情報を色によって区別・分類することは、普段からよく行われており、一番大まかな分類を行う程度の数であれば、対応を覚えておくことは負担ではないと考えられる。

3.3.3 利用フェーズにおけるアプローチ

利用フェーズにおける機能要件に対する、本研究のアプローチの詳細について述べる。利用フェーズにおける機能要件は、見返し行為の支援、関連情報の検索・表示支援、共有の支援の3つである。本研究で提案する手法では、上記の3つの機能要件を、スクラッピングに特化した表示機能、関連情報の自動スクラッピング機能、共有機能、の3つの機能を有するビューアを実現することで満たしている。以下に、それぞれについて説明する。

スクラッピングに特化した表示機能

ビューアは、ユーザが見返しを容易に行えることを目的とし、スクラップされた情報と共に保存されたメタ情報（雑誌名、発行年度及びユーザが独自に付加した情報等）を表示する機能を備える。また、ユーザから付箋を通して伝達された分類方法に基づいて、見やすく分類し、表示する。

関連情報の自動スクラッピング機能

ユーザがスクラッピングした情報の関連情報にすぐにアクセスするために、スクラッピングされた情報からキーワードを抽出し、Web 上から関連情報を検索し、自動的に蓄積する機能をビューアは備える。また、蓄積された関連情報と元のスクラッピング情報との関連を表示する。

共有機能

ユーザがスクラッピングした情報を他のユーザと容易に共有するために、スクラッピング情報共有機能をビューアは備える。共有機能は、ユーザが蓄積したスクラッピング情報と関連のある情報を、本研究の手法を利用する全ユーザのスクラッピング情報の中から検索し、共有が許可された情報を自動的にスクラッピングする機能である。また、関連情報の自動スクラッピング機能と同様に、共有された情報とユーザのスクラッピング情報との関連を表示する。

3.4 本章のまとめ

本章では、まず本研究の対象とする紙発見デジタル蓄積型スクラッピングについて考察し、その形態のスクラッピングを支援するための機能要件を導いた。次に関連研究を調査し、上述した機能要件を満たしたシステムが存在しないことを述べた。そして、本研究のアプローチである印付け動作を用いたスクラッピング手法を述べ、その特徴を説明した。

第4章 Mark itの設計

本章では、印付け動作を用いたスクラッピングシステム Mark it の設計について述べる。設計の概要と Markit を構成する各モジュールについて説明する。

4.1 Mark it

本節では、本研究の提案する印付け動作を用いたスクラッピング手法を実現するシステム Mark it について述べる。Mark it の概要と想定環境を述べ、次にハードウェア構成とソフトウェア構成を説明する。

4.1.1 Mark it の概要

Mark it システムは、紙発見デジタル蓄積型スクラッピングを支援することを目的としたシステムである。紙メディアに対する印付け動作の利用とデジタルメディアの特性を生かしたビューワによって、紙メディアからデジタルメディアへのスクラッピングを、容易かつ直感的に行うことを可能にする。さらにビューワは、スクラッピングされた情報の関連情報を探索、スクラッピングし、関連情報のグルーピングを提供することで、利用フェーズの支援を実現する。Mark it システムは、紙メディアの情報をデジタルメディアとして変換するため、2次元バーコードとカメラを用いる。

4.1.2 想定環境

本システムは、雑誌や新聞などの紙メディア上の記事を対象とする。また想定環境として、記事には各々に対応した2次元バーコードが印刷されており、記事と同一内容のデジタルメディアをデータベースから取得可能である環境とする。さらに、ユーザの上部にはwebカメラが設置されており、ユーザはその下で読書を行うことを想定する。

4.1.3 ハードウェア構成

Mark it のハードウェア構成について説明する。ハードウェアは、Mark it サーバ、ユーザ PC、WEB カメラ、ディスプレイの4つから構成される。ハードウェア構成図を4.1に示す。以下に、それぞれのハードウェアについて説明する。

Mark it サーバ

Mark it サーバは、世界に1つ設置され、全ユーザがアクセスする。雑誌記事の記事内容とメタ情報のデータベースと全ユーザのスクラッピング記事情報のデータベースを保持している。ユーザ PC の要求に従い、記事情報やスクラッピング記事情報を渡す。

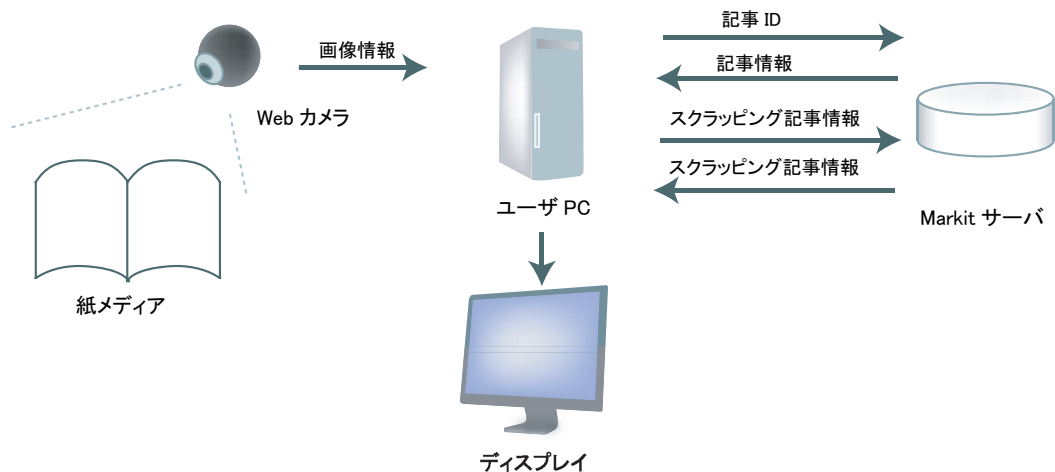


図 4.1: ハードウェア構成図

ユーザ PC

ユーザ PC は、各々のユーザが保持するコンピュータである。スクラッピング記事情報を保持しており、印付け動作の取得や紙メディアからデジタルメディアへの変換、ビューワ管理を行う Mark it クライアントシステムが動作している。Mark it クライアントシステムは、Web カメラから情報を受け取りディスプレイへスクラッピングした記事の情報を渡す。また、ユーザがスクラッピングした記事の情報データベースを保持する。

Web カメラ

Web カメラは、紙メディアの情報をデジタルメディアへ変換する。Web カメラを用いて、印付け動作と 2 次元バーコードを取得することで、スクラッピング対象の記事を取得する。

ディスプレイ

ユーザ PC 上で保持されているスクラッピング記事情報をもとに、Mark it ビューワの表示を行う。

4.1.4 ソフトウェア構成

次に、Mark it のソフトウェア構成について説明する。Mark it のソフトウェアは、Markit 設定リスト、記事情報データベース、スクラッピング記事情報データベース、共有スクラッピング記事情報データベース、発見・蓄積モジュール、ビューワ管理モジュールの大

きく6つから構成される。ソフトウェア構成図を図4.2に示し、各ソフトウェアについて説明する。

Markit 設定リスト

付箋の色と分類の対応リストである。Markit ビューワを用いて、ユーザによって生成される。

記事情報データベース

記事情報データベースには、雑誌に掲載されている記事の内容とメタ情報が蓄積されている。

スクラッピング記事情報データベース

スクラッピング記事情報データベースは、個々のユーザがスクラッピングした記事の情報が蓄積されている。

共有スクラッピング記事情報データベース

共有スクラッピング記事情報データベースには、全ユーザのスクラッピング記事情報が蓄積されている。

発見検知・蓄積モジュール

発見検知・蓄積モジュールは、ユーザのスクラッピング欲求を検知し、蓄積の際に用いる記事の分類情報とユーザが付加した注目箇所を取得する。さらに、記事情報データベースからはスクラッピングする記事のメタ情報を受け取る。それらの情報を用いて、ユーザPC上のスクラッピング記事情報データベースに蓄積する記事情報を作成する。

ビューワ管理モジュール

ビューワ管理モジュールは、Mark it ビューワを管理するモジュールである。関連情報の自動スクラッピングやスクラッピングされた記事情報やメタ情報の表示を行う。

4.2 各モジュールの設計

本節では、上述したソフトウェアに含まれる各データベースとモジュールの設計について詳しく説明する。

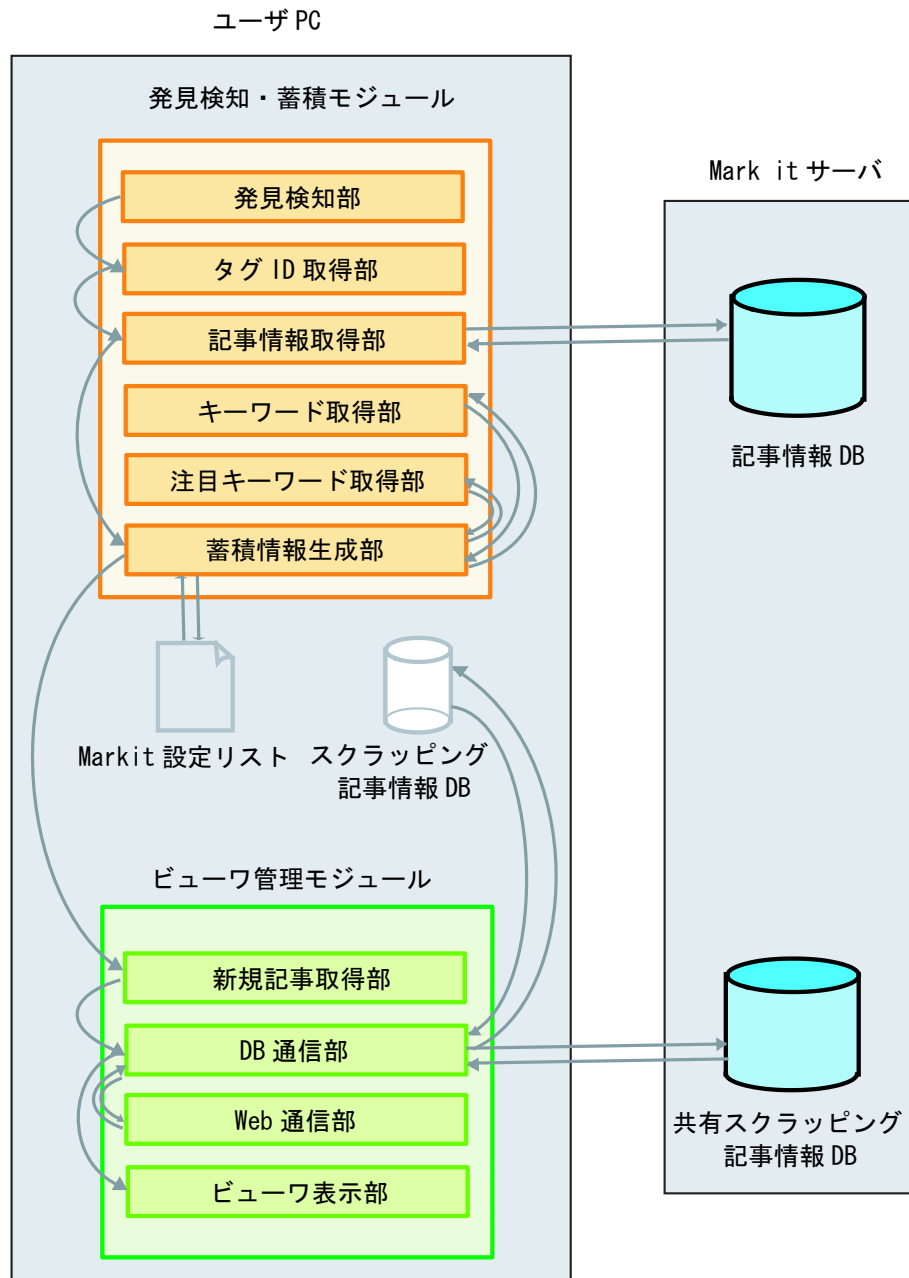


図 4.2: ソフトウェア構成図

4.2.1 記事情報データベース

記事情報データベースは、Mark it サーバ上に設置され、雑誌に掲載されている記事の内容とそのメタ情報を記録するための Article テーブルを保持する。Article テーブルを表 4.1 に示す。

表 4.1: Article テーブル

id	記事の ID
title	タイトル
text	本文
image1	記事内で使われている画像の URL
image2	記事全体の画像の URL
url	雑誌のサイト URL
genre1	記事の広いジャンル
genre2	記事の狭いジャンル
explanation	雑誌名と号数

4.2.2 スクラッピング記事情報データベース

スクラッピング記事情報データベースは、ユーザ PC 上に設置され、2つのテーブルで構成される。1つ目は、システムが自動的にスクラッピングした記事も含む全記事のテーブルである MyArticle テーブルである。2つ目は、ユーザがスクラッピングした記事とその記事のメタ情報の関連を示す MarkedArticle テーブルである。MyArticle テーブルは、Article テーブルと同様に設計する。ただし、記事の ID は各々のユーザが保持する Article テーブルの中でユニークに定められる。MarkedArticle テーブルを、表 4.2 に示す。

表 4.2: MarkedArticle テーブル

id	記事の ID
markdate	スクラッピングした日時
genre	記事の分類
keyword	キーワード
markedKeyword	注目キーワード
related	関連記事の ID

4.2.3 共有スクラッピング記事情報データベース

共有スクラッピング記事情報データベースは，Markit サーバ上に設置され，全ユーザのスクラッピング記事情報のテーブルである AllArticle テーブルを保持する．AllArticle テーブルは，MyArticle テーブルと同様に設計する．

4.2.4 発見検知・蓄積モジュール

発見検知・蓄積情報取得モジュールは，発見検知部と，タグ ID 取得部，キーワード取得部，注目キーワード取得部，記事情報取得部，蓄積情報生成部からなる．図 4.3 にシーケンス図を示す．

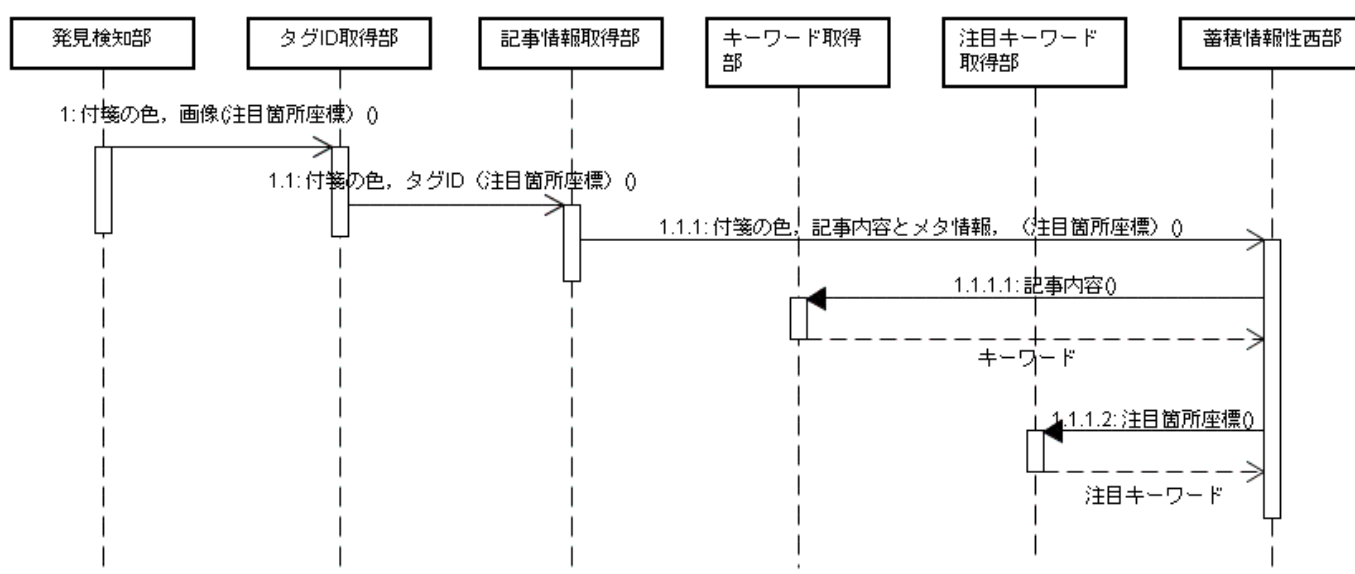


図 4.3: 発見検知・蓄積モジュール シーケンス図

発見検知部

付箋検知部は，常時 Web カメラから画像を取得し，画像解析を行って付箋の貼付を検知する．付箋の貼付を検知すると，その後数秒間ペンによる印付けを待機する．ペンによる印付けが行われた場合は，注目箇所の座標と画像，取得した付箋の色を，ペンによる印付けが行われなかった場合は画像と付箋の色のみを，タグ ID 取得部へ渡す．

付箋の貼付とペンによる印付けは，Web カメラから取得する画像における色の変化によって検知する．本システムで提供する色を数種類定め，定められた色の増加が閾値を超過した場合に，付箋が貼られた，あるいはペンによる印付けがなされたと判断する．ペンによる印付けは，印付けが行われた箇所の最も左上の 2 次元座標と最も右下の 2 次元座標を取得し，整数の 2 次元配列としてタグ ID 取得部へ渡す．

タグ ID 取得部

タグ ID 取得部は、画像から 2 次元バーコードを取得し、2 次元バーコードからタグ ID を取得する。タグ ID を取得した後、付箋の色と注目箇所の座標、タグ ID を記事情報取得部に渡す。

記事情報取得部

記事情報取得部は、記事情報データベースからタグ ID をキーとして記事内容と記事のメタ情報を取得し、付箋の色と注目箇所とともに、蓄積情報生成部に渡す。

蓄積情報生成部

蓄積情報生成部は、Mark it 設定リストを用いて付箋の色から記事を分類すべきジャンルを取得する。また、キーワード取得部に記事情報取得部から受け取った記事内容を渡してキーワードを取得する。さらに、注目キーワード取得部へ注目箇所座標と記事全体の画像を渡し、注目キーワードを取得する。

分類ジャンル、キーワード、注目キーワードの記事のメタ情報に追加し、記事情報に ID を付けてスクラッピング記事情報データベースに蓄積する。また、ビューワ管理モジュール内の新規記事取得部へ、新しくスクラッピングされた記事の ID をソケットを用いて送信する。ソケットの利用は、ビューワの利用先と付箋の貼付を取得する PC が異なる場合にも対応するためである。

キーワード取得部

キーワード取得部は、形態素解析を用いて渡された文字列の中から、キーワードを取得し、カンマ区切りの文字列で返答する。

注目キーワード取得部

注目キーワード取得部は、記事全体の画像から注目箇所部分の画像を抜き出し、OCR を用いて文字情報を取得する。取得した文字情報をキーワード取得部へ渡し、返されたキーワードを注目キーワードとしてカンマ区切りの文字列の形で返答する。

4.2.5 ビューワ管理モジュール

ビューワ管理モジュールは、新規記事取得部、DB 通信部、Web 通信部、ビューワ表示部で構成される。

新規記事取得部

新規記事取得部は、発見検知・蓄積モジュールから新しくスクラッピングされた記事 ID を受け取り、DB 通信部へ渡す。

DB 通信部

DB 通信部は、スクラッピング記事情報データベースとビューワ表示部間のやり取りを行う。新規記事取得部から新規に追加された記事の ID を受け取った場合は、受け取った ID をキーにスクラッピング記事情報データベースに問い合わせ、スクラッピング記事情報を取得する。記事のメタ情報のうちのキーワードと注目キーワード、記事のジャンルを Web 通信部へ渡し、関連情報の URL とタイトルを取得する。また同じように、キーワードと注目キーワードをキーに、共有スクラッピング記事情報データベースへ問い合わせ、他ユーザのスクラッピングした記事の中から、関連のある記事を取得する。これらの関連情報を、元の記事の関連情報としてメタ情報に追加し、スクラッピング記事情報データベースを更新する。最後に、MarkedArticle 及び MyArticle と同様の要素を持つオブジェクトである MarkedArticle オブジェクトと MyArticle オブジェクトを生成し、ビューワ表示部へ渡す。

Web 通信部

渡されたキーワードと記事のジャンルを元に Web 上で検索を行う。記事のジャンルに対応するポータルサイト内での検索と、Web 全体での検索の2種類を行い、サイト内検索結果の上位順、Web 全体での検索結果の上位順の順番に、Web サイトのタイトルと URL を返す。

ビューワ表示部

ビューワ表示部は、DB 通信部から渡された2つのオブジェクトをもとに、スクラッピングした記事の表示を行う。分類ごとの表示と、同一分類内の全スクラッピング記事のリストと個々の記事の詳細表示を行う。また詳細表示では、記事内容と記事のメタ情報の両方を表示する。

4.3 本章のまとめ

本章では、本研究で提案する印付け動作を用いたスクラッピング手法を実現する Mark it システムの設計について述べた。本システムのハードウェア構成とソフトウェア構成を述べ、各モジュールの詳細な設計を説明した。

第5章 Markitの実装

本章では，Markit システムの実装について述べる．まず実装環境について述べ，その後各ソフトウェアの実装について述べる．

5.1 実装環境

本システムは、Markit サーバ及びユーザ PC 用と、ビューワ管理モジュール用の 2 台の PC を用いて実装を行った。実装環境を表 5.1 及び 5.2 に示す。

表 5.1: Markit サーバ及びユーザ PC 実装環境

CPU	Pentium M 1.6GHz
主記憶	1GB
OS	WindowsXP
javaVM	J2SDK 1.5.0
データベース	MySQL
形態素解析ライブラリ	茶筌
2次元バーコードライブラリ	JARToolKit
OCR ライブラリ	Panasonic 活字認識ライブラリ Ver.11
Web カメラ	Logitech Qcam Fusion

表 5.2: ビューワ管理モジュール実装環境

CPU	2GHz Intel Core Duo
主記憶	2GB
OS	MacOSX10.4.7
javaVM	J2SDK 1.5.0
Web 検索ライブラリ	GoogleAPI

データベースには MySQL を使い、形態素解析を行うライブラリとして茶筌 [6] を用いた。2次元バーコードは JARToolkit を用いて実装した。OCR ライブラリは、Panasonic 活字認識ライブラリ Ver.11[7] を用いた。また、Web 検索ライブラリは GoogleAPI[8] を用いた。

5.2 各モジュールの実装

本節では、Mark it システムを実現するソフトウェアの実装について述べる。まずデータベースについて説明し、次にソフトウェアモジュールについて説明する。

5.2.1 データベース

データベースは全て MySQL を用いて実装した。以下に、スクラッピング記事情報データベースの ER 図を示す (図 5.1)。

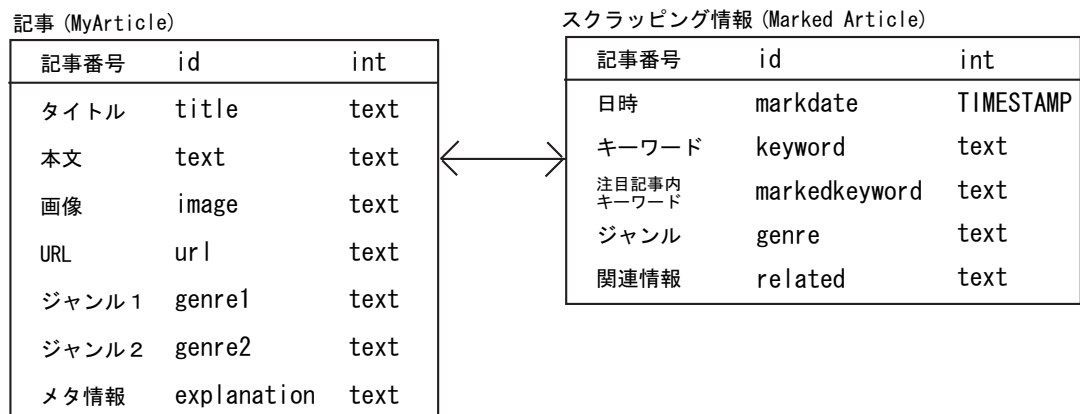


図 5.1: ER 図

MyArticle データベースについて説明する．id は、各記事につけられたユニークな ID である．title と text, image はそれぞれ記事のタイトルと本文、記事内の画像の URL である．url は雑誌の URL であり、genre1,2 はそれぞれおおまかなジャンルと細かいジャンルを示す．explanation は、記事自体のメタ情報であり、掲載雑誌名と号数の 2 つがある．

MarkedArticle テーブルについて説明する．id はスクラッピングした記事にユニークにつける ID である．markdate はスクラッピングした日時、keyword は記事のタイトルと本文を形態素解析して得られたキーワード、markdKeyword は、注目箇所内のキーワードである．genre はユーザが示した分類情報であり、related には関連情報の記事 ID がカンマ区切りで入っている．

5.2.2 発見検知・蓄積モジュール

本項では、発見検知・蓄積モジュールの各部の実装について述べる．

発見検知部

発見検知部は、画像解析を用いて付箋の貼付を取得している．

付箋の貼付は、画像の差分を取得しており、閾値を超える差分が生じた際に、差分発生部分の色と大きさを、付箋の色と大きさと比較することで取得している．また、検知する付箋の色は、ピンク、紫、黄緑、青、オレンジの 5 種類を実装した．図 5.2 に示す．ペンによる印付けには、ピンク色を用いた．図 5.3 に示す．

タグ ID 取得部

タグ ID 取得部は、Web カメラから取得した画像内から 2 次元バーコードを検知し、そのタグ ID を取得する．



図 5.2: 5 種類の付箋の色

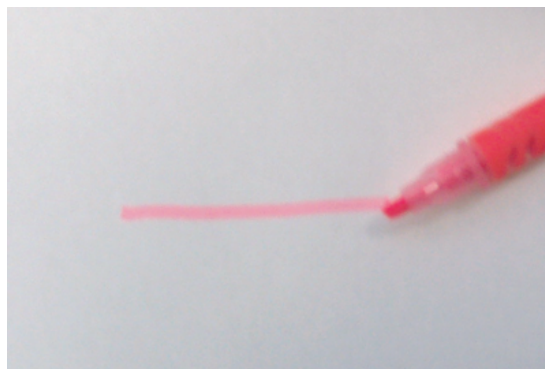


図 5.3: 印付けに用いるペンの色

2次元バーコードには, JARToolKit を用いた. JARToolKit は, ワシントン大学の HITL (Human Interface Technology Lab) によって開発された AR ライブラリである ARTooKit の Java 版ライブラリである [9][10]. 図 5.4 のような黒い四角形の枠を持つ 2 次元バーコードである.

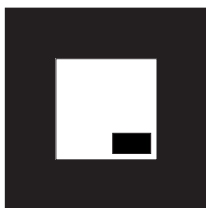


図 5.4: ARToolkit

キーワード取得部

キーワードの取得には形態素解析を用い, 形態素解析ライブラリは茶筌を用いた. また, 形態素解析の結果から未知語と固有名詞をキーワードとした.

5.2.3 ビューワ管理モジュール

本項では, ビューワ管理モジュールの各部の実装について述べる.

Web 通信部

Web 通信部は, スクラッピングされた記事の関連情報を Web から取得する. Web 検索は GoogleAPI を用いて実装した. また, 記事のジャンルに対応したポータルサイトは以下のように定めた.

- 音楽
HMV (<http://www.hmv.co.jp/>)
- 映画
goo 映画 (<http://movie.goo.ne.jp/>)
- その他娯楽
ぴあ (<http://movie.goo.ne.jp/>)
- お洒落
@cosme (<http://www.cosme.net/>)
- IT
CNET Japan (<http://japan.cnet.com/>)
- その他
MSN 毎日 (<http://www.mainichi-msn.co.jp/>)

利用する検索結果のサイト数は、Web 全体とポータルサイトそれぞれ3件とし、検索結果が3件以下であった場合は全ての検索結果を返答する。

ビューワ表示部

以下にビューワのスクリーンショットを示す。

図 5.5 は、ユーザがスクラッピングした情報を中心に閲覧している図である。左のリストは、全てのスクラッピング記事である。リストから一つを選択すると右側で詳細情報を見ることが出来る。詳細表示の際には、スクラッピングした記事の下には雑誌名とスクラッピングした日時を、記事の周辺には関連情報を表示している。

図 5.6 は、自動スクラッピングされた関連記事を中心に閲覧している図である。

ビューワには、利用の際の利便性のために、以下の2つのスクラッピング情報整理機能を実装している。第1に、ビューワはタブ切り替え式になっており、ジャンルごとに表示を切り替えることが出来る。第2に、不必要な記事は棄却することが出来るように、ゴミ箱を用意している。

5.3 本章のまとめ

本章では、Mark it システムの実装について述べた。第1に実装環境を説明し、次に実装の詳細について述べた。次章では、Mark it の評価を行う。

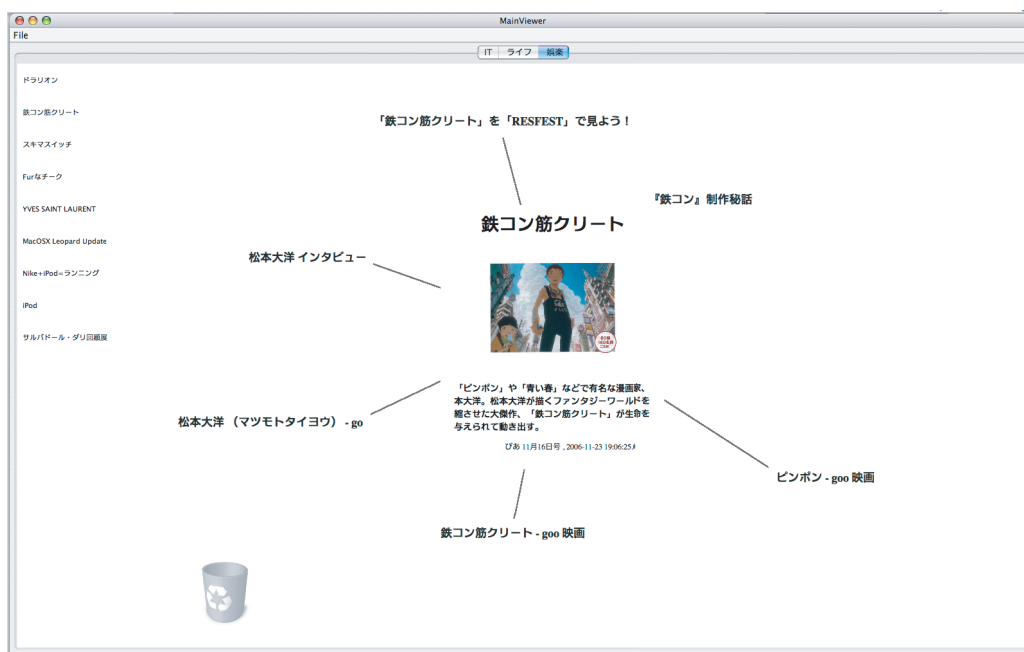


図 5.5: ビューワ スクラッピング記事表示時

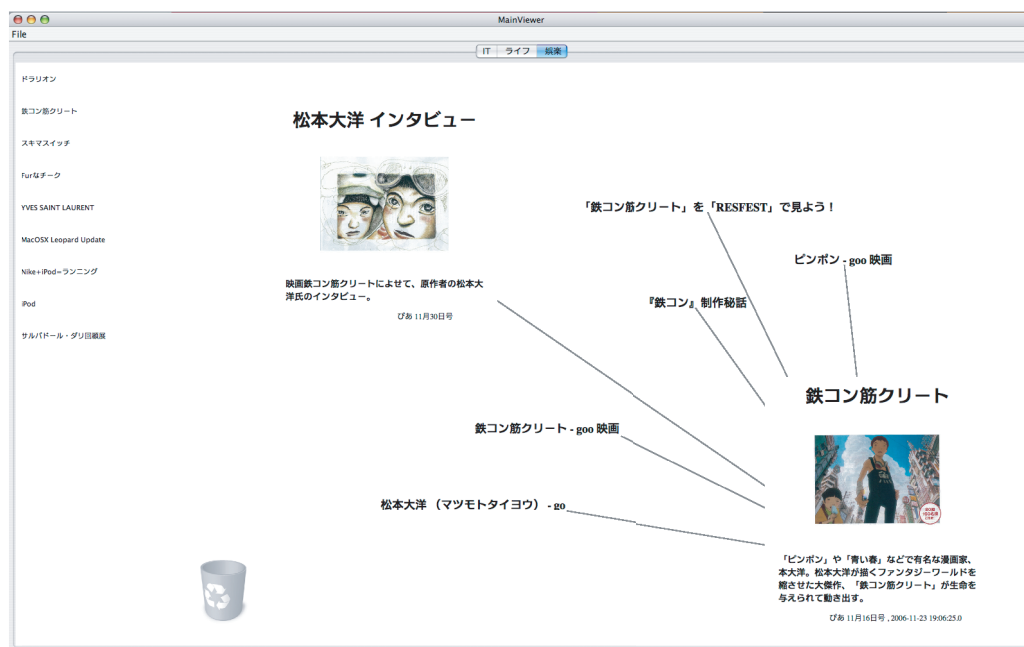


図 5.6: ビューワ 関連記事表示時

第6章 評価

本章では、Markit システムの基本性能評価とユーザビリティ評価を行う。

6.1 基本性能評価

本節では、Mark it の基本性能評価として、付箋貼付の検知にかかる時間と正答率について実験し、考察を行う。

6.1.1 検知所要時間

付箋の貼付を検知するのにかかる時間として、付箋を貼付してから検知された合図の音になるまでの時間を評価した。前章で述べた実装環境において、付箋の検知を20回行い、その平均所要時間を測定する。結果、平均所要時間は2.27(s)となった。測定結果のグラフを図6.1に示す。

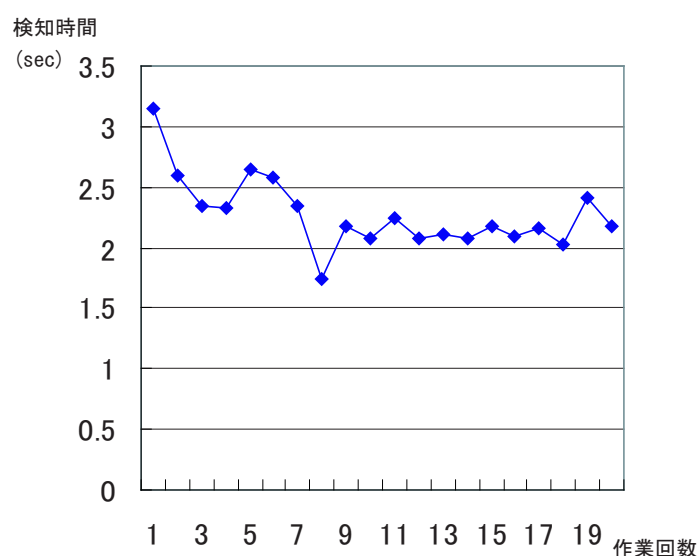


図 6.1: 検知所要時間の測定結果

6.1.2 正答率

付箋の貼付の検知と、貼られた付箋の色の判断における正答率を評価した。2種類の雑誌と3種類の付箋を用いて、各々の組み合わせに対して付箋を真っ直ぐに貼った場合を10回、斜めに張った場合を10回、計120回のスクラッピングを行い、正答率を調査した。付箋は、ピンク、紫、黄緑の3種類を用い、雑誌は背景が白地のものとカラー写真のもの、2種類を用いた。図6.2に示す。

調査の結果、色の判断における正答率は100%であった。貼付の検知における正答率を、表6.1に示す。記事の背景が白地の場合は、93%の高い確率で検知したが、背景がカラー写真の場合は、57%と低確率であった。特に、黄緑の付箋を用いた場合は、30%とほとんど正しく認識しなかった。この結果から、背景色がある場合は認識の精度が落ち、



図 6.2: 評価に用いた雑誌

その中でも背景色と付箋の色の組み合わせによって認識率に差が出ることが分かった。また、付箋の傾きによる正答率の違いは見られなかった。

表 6.1: 貼付検知における正答率

背景	付箋の色	付箋の角度	貼付と色検知					正答率(%)
白地	ピンク	垂直	○	○	○	○	○	100
		斜め	○	○	○	○	○	100
	紫	垂直	×	○	○	○	○	80
		斜め	○	○	○	×	○	80
	黄緑	垂直	○	○	○	○	○	100
		斜め	○	○	○	○	○	100
カラー写真	ピンク	垂直	×	○	○	○	×	60
		斜め	×	○	×	×	○	40
	紫	垂直	○	○	○	○	×	80
		斜め	○	○	○	○	○	100
	黄緑	垂直	×	×	○	○	×	40
		斜め	×	×	×	×	○	20

6.2 ユーザビリティ評価

本節では、ユーザビリティ評価について述べる。ユーザビリティ評価は、Jakob Nielsenによって定義されたユーザビリティの指標である学習しやすさ (Learnability), 効率性 (Efficiency), 記憶しやすさ (Memorability), 間違いにくさ (Error Handling), 主観的満足度 (User Satisfaction) の5つに関して、実験を行い評価した。

6.2.1 実験

Mark it システムにより、紙メディアからデジタルメディアへの変換作業を負担なく行えることを評価するため、次のような実験を行う。

実験環境

実験環境は、図 6.3 のように、机の上に Mark it クライアントが動作する PC とビューワが表示されているディスプレイ及び PC を置き、その前方にデスクトップライトとライトの上にカメラを設置する。ユーザには、ライトの下に置かれた雑誌のスクラッピングを行ってもらう。スクラッピング対象となる記事は 10 種類、付箋の色は 3 種類用意した。



図 6.3: 実験環境

また本実験の被験者は、情報機器の操作に慣れた 10 代から 20 代の学生 10 名である。また、全員が Mark it 初心者である。

実験手順

実験の手順は、想定するシナリオとシステムについての簡単な口頭説明、実験、アンケートの 3 段階で行う。実験は、興味深い記事を発見したところから始まり、付箋を選んで貼り付け、ビューワ上にスクラッピングした記事が表示されたのを確認するまでを 1 連の作業とし、各被験者には 8 回繰り返して作業してもらう。また、付箋を選んで貼り付け、スクラッピングされたことを確認するまでの時間を計測する。実験終了後に、図 6.4 に示したアンケートに記入してもらう。

Mark it システム利用アンケート

性別 男・女 年齢 () 代

このシステムに関する次の質問について、そう思ったかどうかを10段階で答え、該当の数字に○をつけてください。
1＝全くそう思わない 10＝その通りだと思う

付箋を用いた記事のスクラッピングについて

使い方を習得するのは簡単だ	1	—	2	—	3	—	4	—	5	—	6	—	7	—	8	—	9	—	10
作業に負担を感じた	1	—	2	—	3	—	4	—	5	—	6	—	7	—	8	—	9	—	10
必要だと思うことが全て出来る	1	—	2	—	3	—	4	—	5	—	6	—	7	—	8	—	9	—	10
やっていて楽しい	1	—	2	—	3	—	4	—	5	—	6	—	7	—	8	—	9	—	10
付箋の色による分類は分かりやすい	1	—	2	—	3	—	4	—	5	—	6	—	7	—	8	—	9	—	10

ビューワについて

見やすい	1	—	2	—	3	—	4	—	5	—	6	—	7	—	8	—	9	—	10
メタ情報の表示は役立った	1	—	2	—	3	—	4	—	5	—	6	—	7	—	8	—	9	—	10
関連・共有情報は役立った	1	—	2	—	3	—	4	—	5	—	6	—	7	—	8	—	9	—	10
関連・共有情報と元の記事との 関連は分かりやすい	1	—	2	—	3	—	4	—	5	—	6	—	7	—	8	—	9	—	10

感想を自由にご記入ください

実施日 年 月 日

ご協力ありがとうございました

図 6.4: 実験で用いたアンケート

6.2.2 実験結果

実験結果として、一回の作業が終了するまでの平均所要時間を示したグラフを図6.5に、各被験者が作業に要した時間の平均値、最大値、最小値を示したグラフを図6.6に示す。また、アンケートの結果を表6.2に示す。

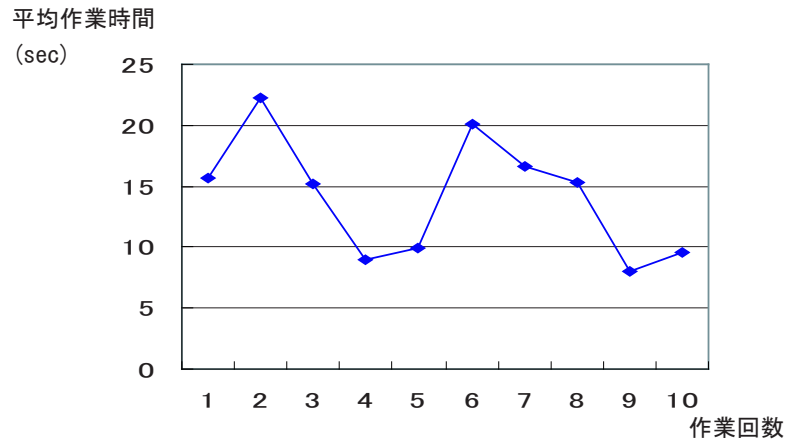


図 6.5: 作業時間平均

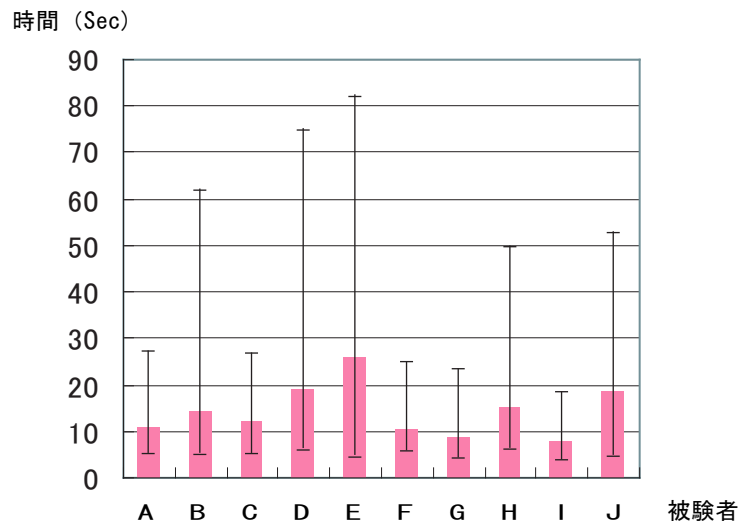


図 6.6: 被験者が作業に要した時間の平均値、最大値、最小値

6.2.3 考察

実験結果から、ユーザビリティの5つの指標に基づいて考察を行う。

表 6.2: アンケートの結果

付箋を用いた記事のスクラッピングについて											
使い方を習得するのは簡単だ	9	10	9	8	8	6	10	10	9	8	8.7
作業に負担を感じた	2	4	8	3	7	5	5	3	5	3	4.5
必要だと思えることが全て出来る	8	6	6	9	5	7	7	5	9	7	6.9
やっていて楽しい	8	7	10	10	7	8	9	8	9	9	8.5
付箋の色による分類は分かりやすい	9	8	8	9	4	9	8	7	6	7	7.5
ビューワについて											
見やすい	9	10	8	10	8	9	7	8	9	8	8.6
メタ情報の表示は役立った	9	10	9	9	8	7	9	9	8	10	8.8
関連・共有情報は役立った	9	9	7	6	9	7	7	7	8	9	7.8
関連・共有情報と元の記事との 関連は分かりやすい	9	9	8	5	4	5	9	7	7	7	7

学習しやすさ

コンピュータスキルの有無に関わらず、全ての被験者が簡単な説明だけで Mark it を利用できた。図 6.2.2 のアンケート結果においても、「使い方を習得するのは簡単だ」という質問に対する 10 人の平均値は 8.7 であり、学習は容易であるといえる。

効率性

図 6.5 のグラフは、ゆるやかな右下がりとなっているが、基本的に作業回数によって作業時間が短くなる傾向は見られず、効率性が高いとは言えない。

記憶しやすさ

Mark it システムは、付箋を貼付するという単純な動作を用いるため、被験者は簡単な説明で作業を理解し、実行することが出来た。よって、時間をおいても忘れにくいと考えられる。

間違いやすさ

図 6.6 の作業時間の最大値から、被験者は最長で 80 秒以上かけて作業を終えたことが分かる。これは、第 6.1.2 項の貼付検知における正答率で述べたように、背景色と付箋の色の組み合わせによってはほとんど検知されないため、成功するまで作業を行ったことが原因である。この実装上の問題は、表 6.2 のアンケート結果の、「作業に負担を感じた」という質問に対する回答の平均値が 4.6 と高くなっていることの原因であると考えられる。自

由記述欄のコメントにおいても、貼付検知の成功率の低さに関するコメントが6件あり、ユーザに負担を与えたことが分かる。今後、付箋の検知部分の実装を見直し、精度を高めることが求められる。

主観満足度

前述のように、作業に対する負担が高いという結果が得られたが、同時に、図 6.2.2 のアンケート結果の「やっていて楽しい」という質問に対する回答の平均値は、8.5 と非常に高い結果が得られた。自由記述欄のコメントにおいても、「付箋に貼るだけで情報を得ることが出来て簡単」「正しく認識したときは快感」というコメントを得た。このことから、今回の実験でユーザが感じた負担は実装面での問題であり、付箋の貼付を用いるという手法自体は好意的に感じられたことが分かる。

また、付箋の色による分類は、平均値 7.5 と良い評価を得られたが、自由記述欄において、「分類が増えるとわかりにくい」というコメントが1件得られた。ビューアに関しては、閲覧性、メタ情報の表示、関連情報の表示に関して、それぞれ高い評価を得た。しかし、「必要だと思うことが全て出来る」という質問に対する回答の平均値は 6.9 であり、自由記述欄のコメントにおいても、「記事のサムネイルが欲しい」「自分の閲覧履歴が見たい」など、ビューアに対する意見が寄せられた。これらの意見を踏まえ、今後の更なる実装が求められる。

6.3 本章のまとめ

本章では、2 種類の基本性能評価と、ユーザビリティ評価を行った。基本性能評価では、付箋の貼付に要する計測時間と、貼付と色の判断の正答率を調査した。ユーザビリティ評価では、付箋の貼付によるスクラッピング作業を 10 人の被験者に対して行い、作業時間を計測し、アンケート調査を行った。次章で、本研究の課題について述べ、まとめる。

第7章 結論

本章では，本論文をまとめる．また，今後の課題と展望を述べる．

7.1 本論文のまとめ

本論文では、印付け動作を用いた紙発見デジタル蓄積型スクラッピングシステムの提案と実装を行った。

紙メディアとスキャナを利用したスクラッピングの場合では、紙メディアをスキャナで読み込んだ後、OCRによる文字情報を取得し、メタデータの付加や分類を行う。この作業は煩雑であり、利用者にとって大きな負担である。

本論文では、このような問題点を解決するデジタルスクラッピングシステムとして**Mark it**を提案した。Mark itでは、紙メディアの記事に付箋を貼ることでデジタルメディアとして蓄積できる。これにより、利用者は興味を持った記事を容易にデジタルメディアとして蓄積できる。Mark itの評価としては、学習しやすさ、効率性、記憶しやすさ、間違いにくさ、主観的満足度の5つに関して、実験を行い評価した。

7.2 今後の課題と展望

7.2.1 付箋の検知における精度の向上

評価の結果から、本システムの付箋検知部分の実装は、検知精度を向上する必要があることが分かった。検知精度が落ちる条件として、記事の背景がカラーである場合に付箋の色によって検知が困難であることが挙げられる。これは、付箋の貼付を検知するために画像解析を行っており、背景がカラー写真である場合は、光の反射やコントラストの変化により、正しく解析が行えないことが原因であると考えられる。

7.2.2 紙メディアからの情報取得

本研究では、紙メディアから情報を取得するための想定環境として、取得する記事と同一内容のデジタルメディアが、取得可能な形で提供されていることを挙げている。そのため、デジタルメディアを提供できないような地方のローカル誌や個人的に作成された情報、過去に既に出版されている書籍などは扱うことが出来ないという問題がある。OCRなどの手法も取り入れ、記事と同一内容のデジタルメディアがない場合でも、紙発見デジタル蓄積型のスクラッピングを可能にする必要がある。

謝辞

本研究を進めるにあたり、御指導を頂きました慶應義塾大学政策・メディア研究科委員長徳田秀幸博士に深く感謝致します。また、貴重な助言を頂きました慶應義塾大学環境情報学部助教授高汐一紀博士に感謝致します。

慶應義塾大学徳田・村井・楠本・中村・高汐・湧川合同研究室の皆様には、折に触れ暖かい御指導と御助言を頂きました。特に、青木俊氏、米沢卓郎氏、山崎俊作氏には、絶えざる励ましとご指導を頂き、深く感謝しています。また、津田恵理子氏、橋爪克也氏をはじめとする HORN グループの皆様、生天目直哉氏をはじめとする KMSF グループの皆様には、多大な協力を頂きました。ここに、感謝の意を表します。

最後に、研究生生活を支えてくれた家族と友人に心からの謝意を表し、謝辞と致します。

平成 19 年 1 月 24 日

川添 瑞木

参考文献

- [1] Catherine C. Marshall and Sara Bly. Sharing encountered information: digital libraries get a social life. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pp. 218–227, New York, NY, USA, 2004. ACM Press.
- [2] 面谷信. 紙への挑戦電子ペーパー 情報世界を変えるメディア.
- [3] Jakob Nielsen. Differences between print design and web design. <http://www.useit.com/alertbox/990124.html>.
- [4] やさしく新聞スクラップ. <http://pac.mediadrive.jp/ynps/>.
- [5] 小林貴訓. 赤外線画像を用いた指先実時間追跡による enhanced desk の実現. ヒューマンインターフェース学会 ヒューマンインターフェースシンポジウム HIS99, pp. 417–422, 1999.
- [6] 茶筌. <http://chasen.naist.jp/hiki/ChaSen/>.
- [7] Panasonic 活字認識ライブラリ ver.11. <http://chasen.naist.jp/hiki/ChaSen/http://panasonic.co.jp/pss/pstc>
- [8] Google api. <http://code.google.com/>.
- [9] jartoolkit. <http://jerry.c-lab.de/jartoolkit/>.
- [10] Artoolkit. <http://www.hitl.washington.edu/artoolkit/>.