# 電子ニュースにおけるダイジェスト機構の実現

3K－3

佐藤円　佐藤理史　篠田陽一
北陸先端科学技術大学院大学 情報科学研究科

## 1　はじめに

現在、電子ニュースを通じて多くの情報が流通し、多くの人々がその情報を利用している。例えば、電子ニュースfjの1カ月の投稿量は、約17000件、約36MBであり、これは、ほぼ新聞の全国紙1紙の1カ月分に相当する分量である[1]。

電子ニュースの一つの特徴は、これらの情報がすべて未編集のまま、読者のもとに届くという点である。現在のニュースリーダーは、ニュースグループ名の一覧や、未読記事数、話題名・投稿者名の到着順表示といった機能を提供する。しかし、従来の情報メディアに見られる、新聞第一面のトピックス表示、雑誌目次、ニュース番組のヘッドラインといった、情報全体を俯瞰し、短時間でそのエッセンスを把握することを可能にしたり、知りたい情報への素早いアクセスを手助けする機能を提供していない。このため、現在の電子ニュースでは、その全体像の把握や目的情報への到達がかなり困難なものとなっている。

このような問題を解決する方法の1つとして、我々は、ダイジェスト機構を提案する[2]。ダイジェストとは、元になる情報の特質をコンパクトにまとめて情報の種類別に整理したものであり、我々が多量の情報に接する際のガイドとして、情報の理解を補助する働きをする。このようなダイジェストがあれば、上記の困難さはかなり解消されると考えられる。電子ニュースの記事は、はじめからオンラインテキストとして存在する。このことは、プログラムによるダイジェストの自動作成にとって非常に有利な特徴となる。

本稿では、fj.meetings を対象として試作したダイジェスト自動作成システム AutoDigest について報告する。

## 2　AutoDigest の概要

AutoDigest は以下の2つのモジュールから構成される。

1. サマリー作成
   各記事からダイジェスト作成に必要な情報（サマリー）を抽出する。
2. ダイジェスト編集
   抽出されたサマリーを編集し、ダイジェストを作成する。

ダイジェストは、HTML 形式で出力され、WWW のブラウザ（例えば、xmosaic）で読むことができる。オリジナル記事の参照は、HTML のアンカーとして埋め込まれたポインタによって実現される。

AutoDigest を実現する主要技術は、1. のサマリー作成である。以下では、これについて述べる。

第4回マルチエージェントと協調計算ワークショップ（MACC'94）

＊＊ 発表・参加募集のご案内 ＊＊

日本ソフトウェア科学会「マルチエージェントと協調計算研究会」（略称 MACC）では、下記要領で第4回ワークショップを開催します。どうぞ奮ってご参加下さい。

1。日時　1994年10月11日（火）、12日（水）、13日（木）（2泊3日）

2。会場

名称: ラフォーレ那須
所在地: 栃木県 那須郡 那須町 湯本 206-959
（中略）

5。論文発表

・論文発表希望者は、A4版2ページ程度のアブストラクトを6部、8月25日まで にプログラム委員長に送付して下さい。
（以下略）

図 1: 記事の例[1]

## 3　サマリー作成

ここでは、日本語記事からのサマリー作成について述べる[2]。図1に、ニュース記事の例を示す。サマリー作成では、このような記事から、記事種別[3]・タイトル・開催期日・開催場所・論文募集締切期日[4]を抽出する。

これらの情報は、記事のスタイル情報および言語表現パターンを手がかりとして抽出する。ここで、スタイル情報による手がかりとは、「タイトルは独立行で表示される」、「開催期日・開催値などの重要情報は、

An Implementation of Automatic Digesting on the NetNews.
SATO Madoka, SHINODA Yoichi, SATO Satoshi.
School of Information Science, Japan Advanced Institute of Science and Technology.
Tatsunokuchi, Nomi, Ishikawa, 923-12, Japan.

[1] HASIDA.94Jun25223514@etlcom.etl.go.jp
[2] 英語記事からのサマリー作成もほぼ同じような方式で実現されている。
[3] 会議告知か、論文募集。
[4] 記事種別が論文募集の時のみ。

表 1: タイトル抽出の手がかり

(a) スタイル情報

- 記事の最上部に置かれることが多い。
- センタリングか左寄せ。
- 記号による飾り (○、★など)、枠などで、会議名称を目立たせる工夫がなされている。

(b) 言語表現パターン

| | タイトルパターン | | | | 通知パターン | |
|---|---|---|---|---|---|---|
| | 接頭辞 | 任意の文字列 | 会議種類名 | 接尾辞 | 1 | 2 |
| 特徴 | (省略可) | | | (省略可) | 単独または重複 | |
| 例 | xx 学会<br>xx 支部<br>第 n 回 | | 研究会<br>フォーラム<br>シンポジウム<br>例会 | '94 | 開催<br>発表募集<br>論文募集 | お知らせ<br>お願い<br>について<br>の件 |
| 凡例 | 第 3 回 | 記号処理 | 研究会 | | 論文募集のお知らせ | |
| | | 著作権 | シンポジウム | | | 参加募集 |

記事種別　論文募集
タイトル　第 4 回マルチエージェントと協調計算ワークショップ (MACC'94)
開催日　941011–941013
開催地　名称:ラフォーレ那須
論文締切　940825

図 2: 日本語会告記事のサマリー抽出実行例

箇条書で表示される」などの、会告記事によく見られる記事の表記形式 (スタイル) から得られる情報である。一方、言語表現パターンによる手がかりとは、「タイトルは、『～研究会』、『～シンポジウム』のようなパターンをとることが多い」、「タイトルの後には『ご案内』などの特徴的な言葉が続く」といった、特有の言語表現から得られる手がかりである。表 1に、タイトル抽出の際に手がかりとなるスタイル情報・言語表現パターンの一部を示す。記事種別・開催期日・開催地・論文募集締切期日の抽出についても、同様の手がかりを利用する。

このような手がかりを用いて、以下の手順に従ってサマリーを抽出する。

1. 各行の行スタイル[5]の特定
2. 記事種別情報の抽出
3. タイトルの抽出
4. 箇条書部分の特定
5. 開催期日・開催地・論文締切の抽出
6. 未発見情報の再探索

図 1からサマリーを抽出した結果を、図 2に示す。

## 4 サマリー抽出の実験

日本語記事のサマリー抽出の精度を調査する実験を、既知データ[6]114 件、および未知データ 97 件に対して行なった。表 2にその結果を示す。

[5]センタリング行・左寄せ行・右寄せ行・タブ行・境界線行。
[6]本システムを作成する際に調査した記事。

表 2: サマリー抽出実験の結果

| | 既知データ | | 未知データ | |
|---|---|---|---|---|
| | 会議告知 (90 記事) | 論文募集 (24 記事) | 会議告知 (73 記事) | 論文募集 (24 記事) |
| | 正解数 (正解率) | 正解数 (正解率) | 正解数 (正解率) | 正解数 (正解率) |
| 記事種別 | 89 (98.9%) | 22 (91.7%) | 70 (95.9%) | 21 (87.5%) |
| タイトル | 84 (93.3%) | 24 (100%) | 63 (86.3%) | 21 (87.5%) |
| 開催期日 | 89 (98.9%) | 24 (100%) | 70 (95.9%) | 20 (83.3%) |
| 開催場所 | 85 (94.4%) | 23 (95.8%) | 70 (95.9%) | 23 (95.8%) |
| 論文締切 | | 22 (91.7%) | | 19 (79.2%) |
| 総合 | 77 (85.6%) | 21 (87.5%) | 57 (78.1%) | 13 (54.2%) |

このように、サマリー抽出は、全体としては、かなり高い正解率を示しており、十分実用に耐えると考えられる。

## 5 おわりに

本稿では、電子ニュースのダイジェストを提案し、その 1 つのプロトタイプとして作成した、fj.meetings のダイジェストの自動作成システムについて述べた。本システムは、現在、JAIST において試験運用しており、WWW によってアクセス可能である[7]。今後の課題としては、サマリー抽出のさらなる精度向上が挙げられる。

## 参考文献

[1] 佐藤円, 篠田陽一. 投稿行動から見た電子ニュース fj. 第 23 回 jus UNIX シンポジウム論文集, pp53-65, 日本 UNIX ユーザー会, 1994. (JAIST Research Report, IS-RR-94-23).

[2] 佐藤円. 電子ニュースにおけるダイジェスト機構の提案と実現. 修士論文, 北陸先端科学技術大学院大学情報科学研究科, 1994.

[7]http://www.jaist.ac.jp/user/sato/nnad/digest-home-j.html