ネットニュース記事群の自動パッケージ化

佐藤 円[†] 佐藤理史[††]

比較的小人数のコミュニティのための電子掲示版としてスタートしたネットニュースは，いまやインターネット上の一大マスメディアとして機能している．しかし，ネットニュースでは，他のテキスト情報マスメディア（新聞，書籍など）のようなユーザの利用目的を考慮した情報のパッケージ化（ある目的に基づき，情報を集め，編集し，1つの製品にすること）が行われていないため，ユーザにとって，いろいろな側面で利用しにくいメディアとなっている．ネットニュースのパッケージ化においては，ネットニュースの特徴である速報性の保持が重要である．そのためには，大量の情報を短いサイクルで編集しなければならず，これには計算機を用いた自動処理による編集の実現が必要である．本論文では，ユーザが，過去の記事群から特定の情報を素早く探し出せるようにすることを目的に，Sunワークステーションを対象とした日本語・英語の質問応答型ニュースグループの記事群を自動編集し，階層的に分類された質問応答集（QA-Pack）を生成する方法について述べる．これを実現する中心技術は，重要文抽出技術と分類コード決定技術であり，情報の表示にはハイパテキストを用いる．

Automated Editing for Packaging Netnews Articles

MADOKA SATO[†] and SATOSHI SATO[††]

The USENET news (netnews) started as electronic bulletin boards for a small community. Now it serves as a mass communication medium on Internet; netnews users are various and so are their objectives. Differently from other mass communication media (e.g. newspaper, magazine), however, there is no information packaging (gathering and editing information, and making it into a product) for netnews; therefore, users become annoyed at netnews when they look for certain information. The automated editing is the approach that solves this problem. In this paper, we discuss the design and implementation of a novel system that edits articles of newsgroups about Sun workstations: the system categorizes articles, structures articles hierarchically, and digests articles all automatically. We overview the significant techniques of this system: summary extraction technique and category code assigning technique. We also describe the representation of the edited information using hypertext. The automated editing system edits a large collection of netnews articles, and generates the package in short time; the system updates the package in short cycles to encompass the latest information.

1. はじめに

比較的小人数のコミュニティのための電子掲示版としてスタートしたネットニュースは，コンピュータネットワークの成長とともにその規模を拡大し，現在では，インターネット上の一大マスコミュニケーション・メディア（以下，マスメディアと略記）として機能するまでに至っている．この成長過程において，「発言したい人はだれでも記事を投稿することができ，その記事はそのまま配送される」という電子掲示版の特徴はそのまま保持されてきており，これが，ネットニュースを他のマスメディアから差別化する大きな特徴となっている．

この特徴は，投稿者にとっては大きなメリットである．なぜなら，投稿の機会はだれにでも平等に開かれており，いつでも好きなときに投稿できるからである．しかしながら，読者の層もその目的も多様になっている現在では，この特徴は読者にとっては，かえって大きなデメリットとなる場合がある．たとえば，ネットニュースは，新しい計算機にあるソフトウェアをインストールしようとしてうまくいかなかった場合に「それに関連する記事が流れていないかどうか調べる」，というような目的でしばしば利用されるが，求める記事を見つけ出すことはそれほど簡単なことではない．これは，ネットニュースにおいては，読者の利用目的を

† 金沢学院大学文学部
Faculty of Literature, Kanazawa Gakuin University
†† 北陸先端科学技術大学院大学情報科学研究科
School of Information Science, Japan Advanced Institute of Science and Technology

考慮したパッケージ化が十分に行われていないことに大きな原因があると考えられる．

既存のテキスト情報マスメディア（新聞，雑誌，書籍など）では，読者（以下，ユーザと呼ぶ）が，そのメディアをどんな目的でどのように利用するかを想定した情報のパッケージ化が必ず行われている．情報のパッケージ化とは，ユーザとその目的を想定し，それに合致するように情報の収集，選択，加工を行うことである．このうち，情報の収集を除いた部分は通常「編集」と呼ばれ，具体的には，分類，選別，組織化，量の調節，書換え，レイアウトデザインと割付け，ナビゲーションのための手がかりの付加，などがこれに含まれる[1]．これらの作業は，readability と legibility☆を向上させるために行われるものである．

たとえば，新聞のパッケージ化は，次のように行われる．読者が新聞を読む目的としては，その日のニュースの項目だけを知りたい，概要を知りたい，詳しく知りたい，特定分野のニュースだけを知りたい，などのいくつかが想定される．新聞を制作する場においては，記者が取材をして事実を集め，それらを基に記事の下書きを作る．記事の下書きは，すべてコピーエディタの元に集められる．コピーエディタは，デザイナ，レイアウトエディタと共同で，ユーザの目的を考慮し，記事にヘッドラインやサマリーを付け加え，全記事の統一性がとれるように記事の文体を書き換え，記事を分類する．そして，同じ分野の記事は同じ紙面内に収め，そのうち重大なニュースは目立つように配置し，さらに，その日のニュースのダイジェストを添え，1つの完成された製品にする．

現在，いくつかのニュースグループを対象に手作業で作成されている FAQ（Frequently Asked Questions）やダイジェストは，記事群を主な情報源として作成されたパッケージ情報である．しかしこれらは，定期的な更新が行われないものや作成が中断されるものが多く☆☆，ネットニュースの重要な特徴である速報性が保持されていない．そのため，たとえばユーザが最新の記事を含んだダイジェストを必要としているときや，つい最近リリースされた新しいソフトウェアに関する情報を探したいときなどには，不十分である．

☆ Readability は，内容の難易度を測る尺度であり，ligibility は，活字が読まれ，理解される際のスピードと正確さを測る尺度である[1]．あえて訳すならば，readability は「分かりやすさ」，legibility は「読みやすさ」となるだろう．本論文では，この2つの言葉を英語のまま使うことにする．

☆☆ これらは，個人のボランティアが手作業で作成しているため，増加し続ける記事数に対応できず，定期的に更新されないことや中断されることが多い[2]．

ユーザが現在のネットニュースに求めているのは，プリントメディアで出版されない情報（特殊情報，隙間情報など）とまだ出版されていない最新情報である場合が多いことから考えると，速報性を保持していないパッケージ化には大きな欠陥があるといえる．

ネットニュースの速報性を保持したパッケージ化を行うためには，記事群の編集を，ごく短いサイクル（たとえば，1日単位など）で繰り返し行わなければならないが，既存の FAQ の例が示すように，これを人手で行うことは現実的ではない．つまり，「速報性」を保持したネットニュース記事群のパッケージ化は，計算機による自動処理によって初めて可能となると考えられる．

我々が先に提案したニュースグループのダイジェスト自動生成システム[3),4)]は，非常に初歩的な段階ながら，ネットニュース記事群の自動パッケージ化を実現したものと考えることができる．しかしながら，これらのシステムが生成するダイジェストは，それぞれのニュース記事の要約情報をある決められた順序に並べたものであり，ユーザと利用目的を想定したパッケージ化，あるいは編集という概念は希薄であった．

本論文では，ユーザが記事群の中から求める情報を素早く探し出せるようにすることを目的とし，Sun ワークステーションを対象とした日本語および英語の質問応答型ニュースグループの記事群を自動的に編集し，質問応答集として利用できるようなパッケージ情報（QA-Pack）にする方法について述べる．計算機による自動パッケージ化の実現には，パッケージ情報のデザインの決定と，自動パッケージ化システムの実現が必要である．以下，まず2章では，パッケージ情報のデザインについて述べ，3章では，自動パッケージ化システムについて述べる．続いて4章では，作成したシステムの評価を行い，5章では，自動パッケージ化に関する議論と関連研究について述べる．

2. QA-Pack のデザイン

本章では，Sun ワークステーションを対象としたニュースグループの記事群をパッケージ化した QA-Pack のデザインについて述べる．対象とするニュースグループは，comp.sys.sun.admin（英語）と fj.sys.sun（日本語）である．これらのニュースグループは，質問応答型ニュースグループであり，Sun ワークステーションに関する質問とそれらに対する応答が数多く投稿される．QA-Pack は，それらを質問応答集として利用しやすいようにパッケージ化したものである．QA-Pack のユーザとしては，専門知識をほとんど持

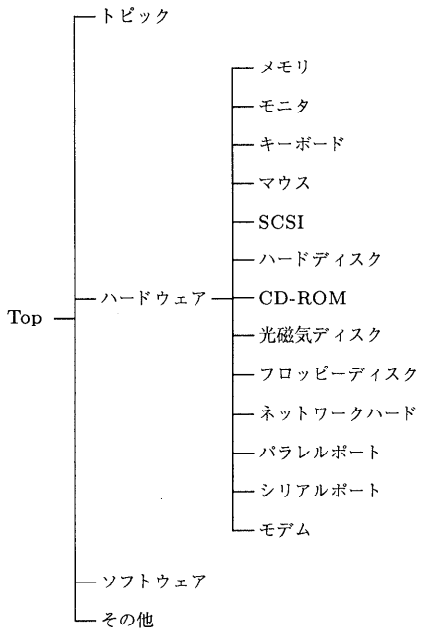


図 1 分類木（一部）
Fig. 1 Concept tree (part).

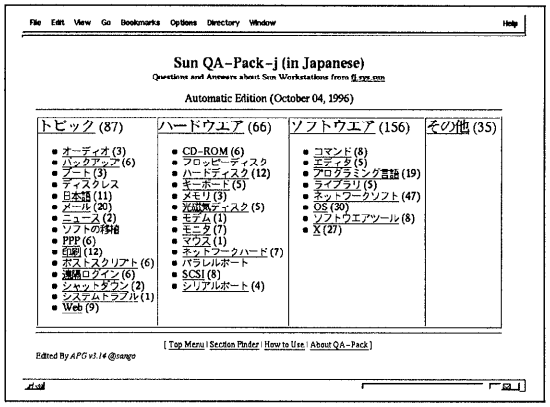


図 2 目次ページ
Fig. 2 Table of contents.

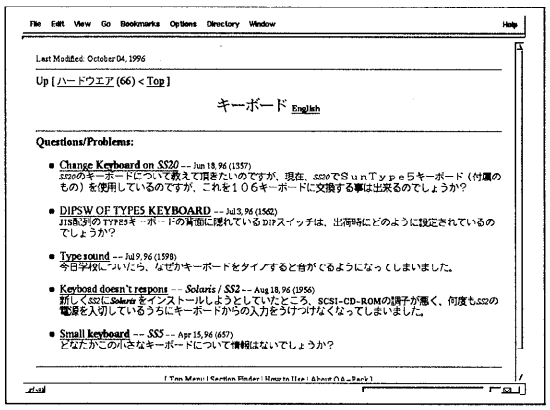


図 3 節点ページ
Fig. 3 A Page for a node.

たないビギナからシステム管理者などのエキスパートまでの，広い層の Sun ワークステーションのユーザを想定する．QA-Pack の目的は，このようなユーザが何か分からないことがあるとき，ネットニュースの記事から答や解決方法を見つけ出すことを支援することである．

2.1 QA-Pack の構造

QA-Pack は，質問応答を階層的に分類したものであり，その基本構造は，分類木（概念木）である．分類木の一部を**図 1** に示す．この分類木において，それぞれの節点は 1 つの概念を表し，その親節点は上位概念，子節点は下位概念を表す．このような基本構造を採用する理由は，後で述べる「目次」式の検索手段を提供するためである．

それぞれの節点は，その節点が表す概念に関連した質問応答集を持つ．この質問応答集に含まれる 1 つの質問応答は，ネットニュースのスレッド（共通サブジェクトを持つ一連の記事群）に対応する．質問応答型ニュースグループにおいては，スレッドの最初の記事が質問記事であり，それ以外の記事が応答記事となる．

2.2 QA-Pack の表示

QA-Pack の表示においては，ハイパテキストを用いる．QA-Pack を構成するページは，QA-Pack の基本構造（分類木）を表示する目次ページ，各節点に付加されたスレッド（質問応答）のリストを表示するページ（節点ページ），各スレッドのより詳しい情報を表示するページ（スレッドページ）の 3 つに分けられる．

目次ページは，QA-Pack の全体像をユーザに提供するものである．**図 2** に目次ページの具体例を示す．

節点ページでは，その節点の位置情報，上位節点および下位節点へのリンクが示され，さらに，その概念に関連したスレッドのダイジェストが列挙される（**図 3**）．このダイジェスト表示は，以下の 2 つの要素から構成される．

(1) ヘッドライン — 質問記事の主題．スレッドページへのハイパリンクを持つ．
(2) サマリー — 質問記事の要点．ヘッドラインを補足する．

スレッドページでは，各スレッドに関するより詳しい情報として，冒頭部に質問記事のヘッドライン，サマリーが表示され，その下に，その質問記事に対する応答記事のリード（応答記事の冒頭部．応答記事本体へのハイパリンクを持つ）が表示される．さらにその

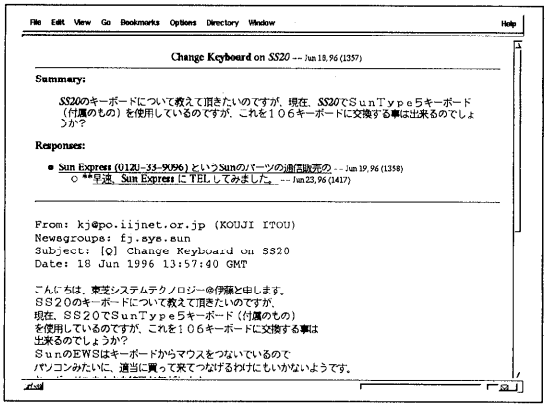


図4 スレッドページ
Fig. 4 A Page for a thread.

下に，質問記事自体が表示される（図4）．

2.3 QA-Pack におけるナビゲーション

QA-Pack の提供する情報検索機能は，電子化テキストの検索において一般的に用いられるキーワードによる全文検索ではなく，目次による検索とキーワードによる概念記述の検索である．また，1つの節点ページ内もしくはスレッドページ内では，多段階ダイジェスト表示が検索を支援する．

2.3.1 キーワードによる全文検索の問題点

電子化されたテキストを検索する手段として一般的に用いられる方法は，テキスト全文を対象とするキーワード検索（以下，キーワードによる全文検索と呼ぶ）である．これは，ユーザが調べたいことを端的に表す単語（キーワード）を入力して，そのキーワードを含むドキュメントを探す方法である．キーワードによる全文検索においては，ユーザが適切なキーワードを思い付くことができれば，素早く求めるものを得ることができる．しかし，的確で，その検索システムに合ったアブストラクション・レベルのキーワードを思い付くことができなければ，答を探し出すのに膨大な手間や時間がかかったり，あるいはまったく答が探し出せないという事態になる．

たとえば，「電話を通してインターネットに接続する方法」を調べる場合，「PPP」や「モデム」といった用語を知っているユーザは素早く答を探し出すことができる．しかし，これらの用語を知らないユーザは途方に暮れることになる．また，Apple Laser WriterII のトラブル対処法について知りたいとき，果たしてその検索システムでは，「プリンタ」，「レーザープリンタ」，「ポストスクリプト・プリンタ」，「Apple Laser WriterII」などのうち，どれが適切なキーワードとなるのかは，実際にいくつかのキーワードを入力して試してみないことには，ユーザは知る由もない．このような例からも分かるように，キーワードによる全文検索は，ユーザが的確なキーワードを思い付くことができるということを前提としており，この前提が満たされない限り有効に機能しない．ビギナを対象ユーザに含める場合，この前提を満たすことは明らかに期待できない．

2.3.2 目次による検索

QA-Pack の提供する中心的な検索手段は，「目次」による検索である．先に述べたように，QA-Pack の目次は，その基本構造となる分類木を表現したもので，ユーザは，この木構造の根節点からスタートし，限られた数の候補（子節点）からの選択を繰り返すことにより，求める情報を探し出すことができる．この方法のメリットは，以下の点にある．

(1) キーワードを知らなくても，求める情報に到達できる可能性がある．なぜならば，限られた数の選択肢の中から1つを選ぶことは，何の手がかりも与えられないままキーワードを思い浮かべることよりも，はるかに容易だからである．
(2) 目次は，いつでもうまくいくナビゲーションシステムである．我々は目次に慣れ親しんでいるので，使い方を間違えることはほとんどない[5]．

この目次による検索により，先に述べたキーワードによる全文検索でうまくいかない例のうち，後者の「Apple Laser WriterII」の例は解決できると考えられる．

2.3.3 キーワードによる概念検索

しかしながら，前者の「電話を通してインターネットに接続する」という例は，目次による検索だけでは解決できない．この例が示す問題の源泉は，ユーザは，その知識レベルによって使う用語が違うという点にある．大まかにいえば，ビギナは日常的な用語を使う傾向が高く（「プリンタがうまく動かない」），エキスパートはより専門的な用語を使う傾向が高い（「printcapの書き方が分からない」）．この問題を完全に解決することは容易ではない．

この問題に対する QA-Pack のアプローチは，キーワードによる概念記述の検索である．まず，分類木のそれぞれの節点（概念）に対して，複数の概念記述を準備しておく．たとえば，概念 <PPP> に対しては，「PPP」，「Point-to-Point Protocol」，「電話によるインターネットへの接続」などを付加しておく．こうして準備された概念記述を，キーワード検索の対象とする．この方法により，ユーザが「PPP」というキーワードを入力した場合は，概念記述「PPP」を通して概念 <PPP> に到達でき，また，「電話」というキーワードを

入力した場合は，概念記述「電話によるインターネットへの接続」を通して概念 <PPP> に到達できることになる．

2.3.4 段階的手がかりの提示

上記の2つの方法は，分類木の中から求める概念を見つけ出すことを支援する機能である．これに対して，求める概念が見つかったのち，その概念に対する質問応答集の中から求めるものを見つけ出すことを支援するのは，多段階ダイジェスト表示である．

先に述べたように，節点ページにおいては，各スレッドのダイジェストが列挙される．この列挙の中から求めるものを見つける第一手がかりとなるのは，ヘッドラインである．ユーザは，これをスキャンすることによって，有望なスレッドの第一次選別を行うことができる．その選別にパスしたもののみ，第二の手がかりであるサマリーを調べればよい．ここで第二次選別が行われる．この選別にもパスした場合のみハイパリンクをたどってスレッドページに進み，スレッドに関するより詳しい情報や質問記事自体を調べればよい．

大まかな情報からより詳細な情報へといくつかの段階を設けて情報を提示することは，従来のメディアでは広く行われていることであり，検索において有効に機能することが期待できる．

3. QA-Pack 自動生成システム

前章で述べたQA-Packを自動的に生成するためには，各スレッドに対して，(1) 質問記事のヘッドラインとサマリー，応答記事のリードを作成すること，(2) そのスレッドがどの概念に関連しているか決定すること，の2つを行う必要がある．この2つが自動生成システムの中心的な処理となる．

3.1 システム構成

QA-Pack 自動生成システムの全体構成を**図5** に示す．本システムは，情報抽出サブシステムとQA-Pack生成サブシステムから構成される．

前者の情報抽出サブシステムは，各記事から，その記事が質問記事の場合はヘッドライン，サマリー，分類コード（分類木の節点に対応する）の3つを，応答記事の場合はリードを生成する．これ以外に，スレッドの認識を可能とするための情報を記事のヘッダ部から抽出する．こうして各記事から生成・抽出された情報は，QAデータベースに蓄えられる．

後者のQA-Pack生成サブシステムは，QAデータベースをハイパテキスト形式のQA-Packに変換する．

本システムは，複数の言語に対応できるアーキテクチャとなっており，1つのシステムで，日本語，英語

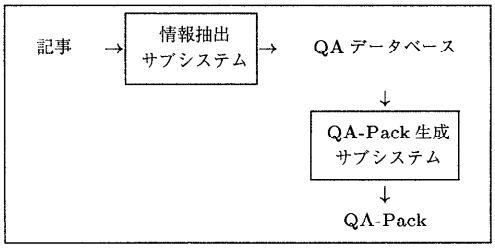


図5 QA-Pack 自動生成システムの構成
Fig. 5 Configuration of automatic QA-Pack generation system.

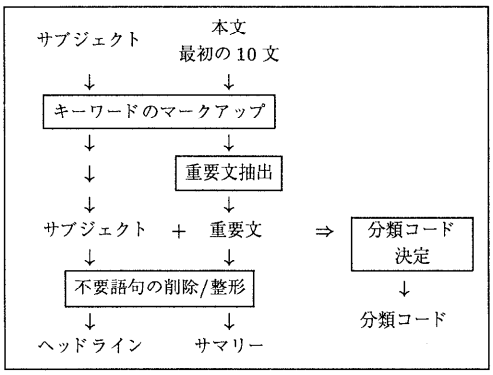


図6 情報抽出処理の概要
Fig. 6 Information extraction processes.

の両言語の記事群を処理することができる．なお，以下の説明の具体例には，日本語記事群に対する処理の例を用いる．

3.2 情報抽出サブシステム

情報抽出サブシステムの主要な処理は，質問記事からのヘッドライン，サマリー，分類コードの生成である．その概要を**図6** に示す．これらの生成は，キーワードのマークアップ，重要文抽出，不要語句の削除/整形，分類コード決定の4つのモジュールによって実現される．

3.2.1 キーワードのマークアップ

本モジュールは，あらかじめ用意された辞書に従ってキーワードを認識し，そこにタグを挿入する．ここでマークアップするキーワードは，分類のために用いる専門用語，OS名，マシン名である．このうち，専門用語に対しては，その専門用語が分類木のどの節点に関係しているかを表す分類コードが辞書中に与えられており，マークアップの際，このコードもタグ中に埋め込まれる．

このマークアップモジュールによって，質問記事のサブジェクト（ヘッダ部から切り出される）と本文の最初の10文のマークアップを行う．**図7** にマークアップの例を示す．

NISの運用下でyppasswdが動かなくなって困っています.

↓

<kwd c=soft.net.nis>NIS</kwd>の運用下で
<kwd c=soft.net.nis>yppasswd</kwd>が動かなくなって困っています.

図7 マークアップ例
Fig. 7 Mark up example.

3.2.2 重要文抽出

本モジュールは，記事の本文の最初の10文からサマリーとして使用する1，2文を抽出する．その方法は，文献4)のサマリー抽出法に基づいている．この方法では，まず特徴的な表現の存在の有無を各文に対して調べ，その結果を総合して最終的に抽出すべき文を決定する．

fj.sys.sunの質問記事は，大別すると，以下の2つのタイプに分類することができる．

I. 特定の事項（ソフトウェアのあるサイト，プリンタのつなぎ方など）について尋ねる記事
II. マシンやソフトウェアを使用するうえでのトラブルや失敗（システムがハングしている，エラーメッセージが出るなど）について述べ，対処法を尋ねる記事

タイプIの記事では，質問者が何を尋ねたいのか，あるいは何をしたいのかを表す文が，その記事において最も重要な文である．一方タイプIIの記事では，質問者がどのような状況に陥っているのかを表す文が，最も重要な文である．そこで，これらの文に特徴的な日本語表現を**表1**のようにまとめ，これらの表現の有無を文字列照合によって判定することによって，重要文を抽出する☆.

3.2.3 不要語句の削除/整形

本モジュールは，サブジェクトと抽出した重要文から，それぞれヘッドラインやサマリーとしてのはたらきに不要な語句を取り除く．以下に，取り除く語句の例を示す.

- サブジェクトから取り除く語句☆☆
  - 修飾文字（例：***）
  - 連続する？や！を1個に（例：??? → ?）
  - 記事目的のサイン（例：[Q]，Wanted:）
  - 省略可能な表現（例：Help with，Question about）
- 重要文から取り除く語句
  - 導入の言葉（例：「さっそくですが」,「さて」,「実は」）
  - 記事の主題に直接関係ない文頭の副詞句（例：「先日より」,「現在,」）

☆ このような方法をとる理由は，ネットニュース記事には，機械処理の際にノイズとなる誤字，脱字，特殊な表現が数多く見られるからである．文字列照合を用いる場合，照合の対象となる部分さえ正しく書かれていれば処理は成功するが，形態素解析を行う場合は，文中に1つでもノイズがあると，解析に失敗する可能性が非常に高くなるからである.

☆☆ 日本語記事のサブジェクトは，ほとんどの場合英語で書かれている.

表1 重要文抽出で用いる特徴
Table 1 Expression patterns for extracting summary.

特徴	品詞	具体例
疑問	（文末）	～か.
	動詞	～について教えて ～についてご存知
	名詞	疑問
	疑問詞	どのようにすれば
探して	動詞	探して
目標	助動詞	～したい，～しよう
作業報告	助動詞	～しているのですが ～中なのですが ～してみた ～したのですが
否定	（文末）	～ん.
	動詞	動かない，できない うまくいかない 分からない，調子が悪い
	助動詞	～してしまう
	名詞	エラー，現象
困惑	動詞	困って

question of swap size on unix

↓

swap size on unix

図8 サブジェクト書換え例
Fig. 8 Subject reformation example.

本モジュールにより，サブジェクトがどのように書き換えられるかの例を，**図8**に示す.

3.2.4 分類コード決定

本モジュールは，サブジェクトおよび重要文からその質問記事の分類コードを決定する．まず，サブジェクトからマークアップされた専門用語（分類コード）を取り出す．もし，専門用語が1つだけ存在した場合は，これを最終的な分類コードとする．もし，専門用語が複数存在した場合は，分類コードごとに数を数え，最も数の多いものを付与する☆☆☆．もし，サブジェクトに専門用語が存在しなかった場合は，重要文に対し

☆☆☆ 出現数が最も多い分類コードが複数ある場合は，それらすべてを選択する．なお，分類コードごとに数を数える際，上位/下位の関係にある2つの分類コードは，下位の（特殊な）分類コードにマージして，数を数える.

て，同様の処理を行う．なお，サブジェクトにも重要文にも専門用語が存在しなかった場合は，「その他」という分類コードを採用する．

3.2.5 その他の処理

上記の説明以外の情報抽出サブシステムの処理には，以下のものがある．

(1) 質問記事からのOS名，マシン名の抽出
OS名，マシン名は，マークアップモジュールでマークアップされる．ここでは，サブジェクト，重要文，その他の本文，の順で探索し，OS名，マシン名を抽出する．

(2) 応答記事からのリードの抽出
応答に関して意味のある最初の1行を切り出す．具体的には，挨拶や引用をスキップし，これら以外の最初の1行をリードとする．

(3) 記事のヘッダから情報抽出
記事ID（ID:），投稿日（Date:），投稿者（From:），サブジェクト（Subject:），リファレンス（Rererence:）の各項目を切り出す．

3.3 QA-Pack生成サブシステム

QA-Pack生成サブシステムは，情報抽出サブシステムによって生成されたQAデータベースを，HTML形式のQA-Packに変換することを行う．その主要な処理は，次の3つである．

(1) スレッドの認識
記事の参照関係を解析し，記事群をスレッド群に変換する．この処理は，記事のヘッダから抽出した，記事IDとリファレンスを用いて行う．

(2) 代用ヘッドラインの作成
記事のサブジェクトがたとえば“Please help”のような場合は，ヘッドラインとしては利用できない（不要語句削除/整形モジュールによって削除される）．このような場合は，分類コード決定に用いられた専門用語をヘッドラインとして採用する．

(3) OS名，マシン名の標準化
たとえば，SparcStation 10は，SS10，SS-10，Sparc-10など各種に記述されるが，これらを標準形式（SS10）に変換する．

(4) 専門用語の強調
専門用語タグをHTMLの強調タグ（）に変換する．

3.4 複数言語への対応

本システムにおいて対象言語に依存する部分は，以下の3つである．

- 専門用語辞書☆
- 重要文抽出で用いる特徴表現（表1）
- 削除する不要語句リスト（3.2.3項）

本システムは，これらの部分を書き換えることによって，他の言語にも適応可能である．

4. 実験と評価

ここでは，本システムの実現を可能にした2つの重要な技術，すなわち重要文抽出と分類コード決定について，その能力を評価する実験を行った．実験は，日本語記事を対象とし，fj.sys.sunから収集した，1）重要文抽出モジュール作成時に調査を行った記事226件（以下，既知データと呼ぶ），2）新しく入手した記事93件（以下，未知データと呼ぶ），の2種類のデータに対し，重要文抽出の精度と分類コード決定の精度を調べた．

4.1 重要文抽出の評価

ここでは，システムが抽出した重要文（サマリー）を，以下の4つのカテゴリーに分類した．

- GOOD：抽出された文はサマリーとして適切である．
- OK：抽出された文は，サマリーとして適切ではないが許容できる．
- NG：抽出された文はサマリーとして不適切である．
- FAIL：抽出できない．

実験の結果を，**表2**に示す．

GOODとOKを合わせると，既知データに対しては83.2%となっており，未知データ（ブラインドテスト）に対しては82.8%となっている．これは，ネットニュース記事のテキストのクオリティが低い☆☆ことを考慮すると，非常に高い精度で重要文を抽出しているということができる．なお，重要文抽出失敗（NGおよびFAIL）の，主な原因は，以下のとおりである．

(1) 質問内容を端的に表す単文（または，2，3文）が，記事中に存在しない．

(2) 重要文に特徴的な表現が存在しない（非常に特殊な形で書かれている）．

表2の右の欄（修正後）は，未知データに対する実験結果をシステムにフィードバックした後のシステムの性能を示している．この表が示すように，重要文抽出においては，システムの性能にそれほど変化がなかった．

☆ 日本語においても，英文表記の専門用語が使われるため，実際には，1つの辞書（日本語と英語の両方の専門用語が混在したもの）を用いている．

☆☆ たとえば，誤字・脱字が多い，文法的誤りが多い，パソコン通信によく見られる特殊な表現が混ざっていることがある，など．

表2 重要文抽出の実験結果
Table 2 Results of summary extraction tests.

	既知データ		未知データ			
			ブラインド		修正後	
GOOD	181	80.1%	67	72.0%	72	77.4%
OK	7	3.1%	10	10.8%	8	8.6%
NG	29	12.8%	12	12.9%	9	9.7%
FAIL	9	4.0%	4	4.3%	4	4.3%
Total	226	100%	93	100%	93	100%

表3 分類コード決定の実験結果
Table 3 Results of classification tests.

	既知データ		未知データ			
			ブラインド		修正後	
GOOD	192	85.0%	60	64.5%	76	81.7%
NG	11	4.9%	11	11.8%	6	6.5%
FAIL	23	10.2%	22	23.7%	11	11.8%
Total	226	100%	93	100%	93	100%

4.2 分類コード決定の評価

ここでは，システムの分類コード付与の結果を，以下の3つのカテゴリーに分類した．

- GOOD：付与された分類コードは適切である☆．
- NG：付与された分類コードは不適切である．
- FAIL：記事に分類コードを付与できなかった（「その他」に分類された）．

実験の結果を，**表3** に示す．

既知データに対しては，85%という高い精度で正しい分類コードを付与することができたが，未知データ（ブラインドテスト）に対しては，GOODが64.5%と，約20%精度が落ちている．これは，専門用語辞書の語彙数の不足が大きな原因である．実際，専門用語辞書（509語）に，未知データから得られた専門用語14語を新たに追加した結果，システムの性能は81.7%まで向上した．しかし，現在の専門用語辞書の語彙数は約500強であり，まだ十分とはいえない．今後専門用語辞書をより拡充することにより，恒常的に高い精度で分類コードを付与することができると考えられる．なお，辞書拡充後の分類コード付与の失敗（NGおよびFAIL）の主な原因は，サブジェクトに質問内容を適切に表すキーワードが存在せず，かつ，適切な重要文を抽出できないため，その結果として適切な分類コードが付与できないというものである．

☆ 複数の分類コードが付与された場合は，それらすべてが適切であることを条件とした．

5. 議論と関連研究

5.1 議　　論

情報を自動的に編集し，パッケージ情報にまとめる研究は，今までにない新しい試みである．これを実現する編集は，これまで「人」が行うものとされ，それを支援するシステム（編集支援システム）は開発されてきたが，編集そのものを行うシステムはほとんど研究されてこなかった．本システムは，人間の編集能力をすべて実現したものではないが，計算機による自動編集の可能性を示した点で大きな意義があると考える．

計算機による自動編集のメリットは次の3つである．

(1) ほとんど処理コストがかからないこと．
(2) 短時間で処理を行えること．
(3) 機械的な処理ならではの統一性が実現できること．

これらの特徴により，短いサイクルでのアップデートや，大量の情報の処理，統一的な方針のもとでの長期的な編集が可能となる．

本システムが行う自動編集の中心的な内容は，情報の分類とナビゲーションのための手がかり付加である．このうち，前者は，質問記事をその内容に従って分類することであり，キーワードを用いたテキスト分類（分類コード決定）によってこれを実現した．また，後者には，分類結果に基づく目次の作成，各スレッドに対するダイジェスト表示などが含まれるが，このうち，ダイジェスト表示のために必要なダイジェスト作成は，特徴的な表現に基づく重要文抽出によって実現した．

本システムにおいて，ニュースグループに依存する部分は，分類木，マークアップのための専門用語辞書，重要文抽出のための特徴的な表現パターンの3つである．本システムを他のニュースグループに適用する場合は，これらの部分をそのニュースグループに適したものに変更すればよい．特に，Sun以外のコンピュータ・システムに関するニュースグループ（fj.sys.*またはcomp.sys.*など）に適用する場合は，専門用語辞書の一部を修正することで対応できると考えられる．

本システムで作成した質問応答集（QA-Pack）は，既存のFAQと比較すると，短いサイクルでの更新が可能なのでつねに最新の情報を取り込める．より詳細に分類されている，行われた議論を再現できる，などのメリットがある．しかし，応答記事の中のどの答が正しいのかの判断（authentication）がなされていないこと，情報源がネットニュース記事に限定されていることなどのデメリットがある．

今後の課題としては，以下のような，より高度な編集機能の実現があげられる．

(1) 統一的なヘッドラインの作成
ユーザが情報を選択する際には，すべてのヘッドラインが，同じ文形（節，句，S+V など）に統一されている方が見やすいと考えられる．今後は，たとえば英語の場合，ヘッドラインにはつねに動詞があることが望ましいとされていることに基づき，すべてのヘッドラインを S+V の形に書き換える方法の研究が必要である．

(2) ヘッドラインとサマリーの連係
サマリーは，ヘッドラインを補完する働きをするので，ヘッドラインに書かれていない情報のみから構成されるべきである．今後は，どのようなヘッドラインを作成したかにより，どのようなサマリーを作成するかを決める方法の研究が必要である．

(3) 繰り返しなされる質問（Frequently Asked Questions）の認識
多くの人から繰り返される質問を認識し，他の質問と差別化する機能を実現する方法の研究が必要である．

なお，既存の FAQ☆を，自動パッケージ化システムの情報源として利用し，QA-Pack に統合するという方法も考えられる．しかし，自動パッケージ化システムは，元情報を独自の方法で分類，選別し，必要であれば個々の文を改変したり，新しい情報を付け加えるものである．既存の FAQ は，その作者の著作物であるため，作者の許可なしに上記の作業を行うことはできない．よって，本研究の段階では，既存 FAQ と自動パッケージ化システムによる QA-Pack との統合の可能性を示すだけにとどめる．

5.2 関連研究

この研究に先立ち，我々は，アナウンス型ニュースグループのダイジェストを自動生成するシステムを作成した[3),4)]．その経験から，元情報の内容をコンパクトにまとめたものを羅列するだけではダイジェストとしては十分ではなく，元情報の適切な分類と，明確なデザインのもとの表示がキーポイントであることを学んだ．分類や明確なデザインに基づいた表示は，既存テキストマスメディア界における編集作業の一部に相当する．このことから，もし，より高度な編集作業を自動化することが可能ならば，ネットニュース記事のように電子化されたテキストを，よりユーザにとって使いやすいパッケージ情報の形に自動的にまとめられるのではないかと考えたのが，この研究の発端である．この意味で，ダイジェスト自動生成システムの研究は，自動パッケージ化の研究の先鞭をつけるものであった．

ネットニュースの自動パッケージ化システムは，ユーザがネットニュースを楽に，かつ有効に利用するためにユーザを補助することを目的とするシステムである．本研究で作成したシステムと同様の目的を持つシステムで，現在実用化されているものには，SIFT[6)]などのネットニュース記事のフィルタリングシステムがある．記事群を組織化し，ユーザにとって分かりやすいようにまとめる自動パッケージ化システムと異なり，これらのフィルタリングシステムは，記事群には手を加えず，キーワード全文検索を用いて記事群をふるいにかけ，ユーザの関心に関係した記事だけを選び出し，それをそのままユーザに提供する方法をとっている．自動パッケージ化システムは，フィルタリングシステムとの連係も可能である．たとえば，ユーザの関心に基づいてフィルタリングされた後の記事群をパッケージ化したり，フィルタリングを行うエージェントがパッケージ化した情報にアクセスするなどの可能性が考えられる．

FAQ をユーザの情報探索に役立てようという研究には，FAQFinder[7)]がある．FAQFinder は，ユーザの質問に，既存の FAQ を利用して答えるシステムである．自動パッケージ化システムと FAQFinder は，両者とも FAQ というパッケージ情報を用いてユーザの情報探索を支援しようとするシステムである．しかし，自動パッケージ化システムは，FAQ 自体を機械的に作成するシステムであるのに対し，FAQFinder は，既存の FAQ を情報源としてユーザによる情報探索を機械的に支援するシステムであり，アプローチはまったく異なる．

HyperNews[8)]は，WWW とネットニュースの融合を目指した研究である．この研究では，投稿されたニュース記事を NNTP を用いて各サイトに配送するのではなく，投稿者が特定の WWW サイト上に記事のありか（URL）を投稿し，読者はそのリンクをたどることによって記事を読むシステムが作成されている．HyperNews と QA-Pack は，読者にとっては，オンライン上で行われた議論が再現されているという点では同じである．しかし，HyperNews では，投稿された記事群を自動的に編集するという機能は実現されていない．

☆ 人手で作成され，ネットニュースに投稿されたり，WWW で公開されたりしているもの．

6. おわりに

本研究では，ネットニュースの記事群を自動的にパッケージ化する方法を提案し，Sun ワークステーションを対象とした質問応答型ニュースグループの記事群を自動的に編集し，質問応答集として利用できるようなパッケージ情報（QA-Pack）にするシステムを実現した．

人間が日常行っているテキストの処理は，大きく，「読む」,「書く」,「翻訳する」,「編集する」の4つに分けられると考えられる．これらのうち，自然言語処理研究が今まで対象にしたきたものは，読む（文章理解），書く（文章生成），翻訳（機械翻訳）であり，編集については，人間が行う編集の支援システムは研究されてきたが，編集そのもののを機械的に実現する研究は行われてこなかった．本研究は，この編集を自動的に行うという新しい試みであるといえる．

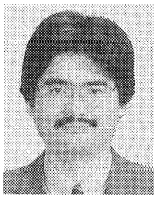
参考文献

1) Moen, D.: *Newspaper Layout Design*, third edition, Iowa State Univesity Press (1995).
2) 奥乃 博，磯崎秀樹，佐藤理史，佐藤 円，山崎憲一：フロー型情報からストック型情報への変換とその可視化の試み，日本ソフトウェア科学会第13回大会論文集，pp.217–220 (1996).
3) 佐藤 円，佐藤理史，篠田陽一：電子ニュースのダイジェスト自動生成，情報処理学会論文誌，Vol.36, No.10, pp.2371–2379 (1995).
4) 佐藤理史，佐藤 円：ネットニュースグループ fj.wanted のダイジェスト自動生成，自然言語処理，Vol.3, No.2, pp.19–32 (1996).
5) Mok, C.: *Designing Bussiness*, Adobe Press (1996).
6) Yan, T.W. and Garcia-Molina, H.: SIFT – A Tool for Wide-area Information Dissemination, *Proc. USENIX Winter 1995 Technical Conference*, New Orleans, LA, USA, pp.177–186 (1995).
7) Hammond, K., Burke, R. and Schmitt, K.: FAQ finder: A Case-Based Approach to Knowledge Navigation, *Proc. 11th Conference on Artificial Intelligence for Applications*, Los Angeles, CA, USA, pp.80–86 (1995).
8) LaLiberte, D. and Braverman, A.: A protocol for Scalable Group and Public Annotations, *Computer Networks ISDN Systems*, Vol.26, No.6, pp.911 918 (1995).

(平成8年10月21日受付)
(平成9年 4 月 3 日採録)

佐藤 円（正会員）

1986年慶應義塾大学法学部政治学科卒業．同年，(株) 総合ビジョン入社．1990年（株）電通総研勤務．1997年北陸先端科学技術大学院大学情報科学研究科博士課程修了．現在，金沢学院大学文学部講師．計算機ネットワーク上のマスコミュニケーション，計算機による情報の自動編集等に興味をもっている．



佐藤 理史（正会員）

1983年京都大学工学部電気工学第二学科卒業．1988年同大学院博士課程研究指導認定退学．同年，京都大学工学部助手．1992年より北陸先端科学技術大学院大学情報科学研究科助教授．京都大学博士（工学）．自然言語処理，機械学習，超並列人工知能などの研究に従事．