知的ニュースリーダ -HISHO- の開発

小作 浩美 内元 清貴 井佐原 均
郵政省 通信総合研究所 関西先端研究センター
{romi,uchimoto,isahara}@crl.go.jp

インターネットの普及に伴い, ネットワークニュースの記事量が増加している. その中において, より効率的なニュース利用を目指し, 新しいフィルタリング技術や検索方法の提案がなされてきている. 我々は対話型ニュースグループを対象にしてユーザの興味にあった1つの記事を中心に同じ話題の記事, あるいは文脈上継りがある記事を抽出するような知的ニュースリーダ -HISHO- (Helpful Information Selection by Hunting On-line) を開発している.
本稿では, 知的ニュースリーダ -HISHO- の概要を紹介し, 関連記事を収集するツールについて説明する. さらに, その評価結果を報告する.

Development of Intelligent Network News Reader -HISHO-

Hiromi OZAKU, Kiyotaka UCHIMOTO and Hitoshi ISAHARA
Communications Research Laboratory, M.P.T.

As the Internet has become very popular, the number of articles generated everyday is increasing rapidly. To make good use of the network news, much research has been done on extracting information from it.
And we have been developing the Intelligent Network News Reader - HISHO -. The system attempts to extract news articles of interest to a user. In this paper, we outline the HISHO system, propose new collection methods for one part of the system, and evaluate them.

1 はじめに

インターネットにおけるネットワークニュースは, 重要な情報源の一つである. 日本では,1994 年ころからインターネットの導入が進み, その利用者も急増してきている. それにあわせて, 流れる情報量も増え, 個々のユーザが本当に必要としている情報が見つけにくくなってきている [1]. ネットワークニュースにおいては, fj のニュースグループだけでも 300 を越すグループが存在し, その投稿数も急激に増加し, かなりの量が流通している.

記事はある程度, 内容に応じたグループに投稿されるが, 利用者の多様化に伴い, 複数のニュースグループ間を移動する話題が存在したり, 適切なグループに投稿されない記事も存在する. 従って, 適切なグループにアクセスしても必要な情報がそのグループに存在する保証がない. よって, 関係のありそうないくつかのグループを結局見なくてはならないことになる.

もちろん, 電子図書館の発達や新聞記事の電子化も進むなか, 情報検索のための研究はいろいろなされており, 実用化もすすんでいる [2]. しかし, ネットワークニュースには, アナウンス型 (新聞記事型) のニュースグループと, 対話型 (討論型) のニュースグループが存在しており [3], アナウンス型のグループにのみ対応するシステム [4] がほとんどである. 対話型ニュースグループは, 話し言葉が利用され, 実際の対話のような記述がなされる. さらにこの種のグループには編集者が存在しないため, 記事に現れる用語や表記方法が一定ではない. そのため, 従来の情報検索技術は対話型ニュースグループの記事検索にはそのまま応用できない. また, まだ読んでいない記事群から検索に必要な適切なキーワードを選択決定することも難しい.

そこで, 我々は, ネットワークニュースの情報をより効率よく利用するため, 知的ニュースリーダ -HISHO- (Helpful Information Selection by Hunting On-line) の提案 [5] を行なってきた. このシステムは, 対話型ニュースグループの記事について, キーワード入力の代わりに興味のある記事を入力とし, その記事につながる話題の流れを追い, ユーザの興味を持った話題に関連する記事群をニュースグループに関係なく抽出することを目的として構築されている. 平成9年度,HISHO システムは JAVA 言語を利用して, アプレット版が完成した. 現在, それを改良してアプリケーション版を作成中である. システム構築に際して, ユーザに示すべき記事を決定する類似記事群の収集ツールの評価実験を行なった. 本稿では, それらについて報告する.

2 知的ニュースリーダ -HISHO-

本システムの利用法としては次のようなものを想定している.

ユーザは多忙で毎日ニュースを読むことはできない. 少し時間を見つけて, 未読記事の中から, いくつかの最新のニュース記事を読む. その中に興味のある記事を見つけ, そ

の記事の話題を理解するために関係する記事群をすべて読みたいと考えた.

このような状況で利用できるニュースリーダとしてHISHO システムは構築されている.

システムの実際の動作手順は, 以下のようになる.

1. ユーザが記事群中に興味のある記事を見つける.
2. 記事のヘッダ部分の情報を元にリファレンスツリーを生成する.
3. リファレンスツリー同士の関連の度合を判定する.
4. ユーザの興味に応じて, さらに関連のあるツリーを収集し, 連結する.

このシステムの主要な特徴は, ユーザの興味に沿って動的に検索することにある. 本システムは辞書を用いずに記事間の意味的距離を測定するため, 特定のニュースグループに範囲を限らずに必要な情報を抽出可能である.

2.1 ニュース記事とリファレンスツリー

ネットワークニュースの記事は, ヘッダとメッセージの2つの部分から構成されている [6]. ヘッダはいくつかのフィールドから構成される. その中に記事の識別子 Message-ID と, 関連する記事の Message-ID からなる References, 記事のタイトルを記述する Subject, 投稿者のアドレスが記述される From, 投稿されたニュースグループを示す Newsgroups, 投稿された時間 Date などがある. Message-ID, References と Subject は記事同士の関係を示している.

また, 賛成反対や冗談の一行だけしかないような情報の非常に少ない記事がたくさん存在する. このような記事を個別に詳細にチェックすることはいろいろな点で無駄が多くなってしまう. References は, ユーザがある記事に対して返答を行なった時には自動的に引用された記事の Message-ID が付くことが多い. そのため, これらのフィールドを利用すればある程度, 関係した記事群を自動抽出することができる. ある程度関係している記事群を調査することで, より的確なタームを抽出することが可能となり, 精度も上がる. References を利用した記事同士の関係はツリー構造になる. HISHO システムでは, この記事群をリファレンスツリーと呼び, 検索等の最小単位とする.

リファレンスツリーは基本的には何らかの話題提起の記事から始まり, それに回答またはコメントする記事で構成される. もし, 全ての記事がニュースサーバ上に存在しているならば, ツリーのトップ記事, 話題提起の記事には References フィールドが存在しないことになる. しかしながら, 実際のニュースサーバにおいて, ある1つのニュースグループ内で, リファレンスツリーを作成し, トップ記事を調べると, References フィールドを持つものが存在した. これは他のニュースグループから移動してきた話題であったために参照元の記事がこのグループ中に存在しない場合や, 参照している記事が何らかの理由でニュースサーバに届かなかった場合, expire により参照記事がサーバ上に存在しない場合などである.

また, 一般にある一定の話題のまとめの記事は, もともとの話題のリファレンスツリーにつながることが少ない. これは, まとめ記事は話題に関係する記事が出尽くしてから投稿されるため, 投稿されるまでに一定の期間, 時間が空いてしまい, その時差のために元記事を見つけるのが困難であり, 元記事を参照せずに投稿してしまうことなどが原因と考えられる. しかし, このようなまとめの記事ほど, ユーザにとって有益であることが多い. また, まとめ記事を理解するために元記事のスレッドが必要であることが多い. そこで, 何らかの理由でわかれてしまったリファレンスツリーの関係を修復するために類似リファレンスツリーの収集を行なう.

一方, ツリーによっては長期間存続し, たくさんの記事が含まれる. そのような場合, 含まれる話題が複数存在することがあり, ユーザの興味に合わない話題については, 表示を避ける処理を行なう必要がある. そのためには, 各話題毎に分離ポイントを把握しておくことも不可欠である. 次にそれらの計算手法について簡単に述べる.

2.2 関連度計算

関連度計算は2つのフェーズで行なわれる. 話題転換点の決定と類似話題の収集である. この計算においては, ヘッダ部分の情報も利用するが, ヘッダ部分の情報だけでは曖昧な点が多く, 完全な結果を出すことはできない. そこで, メッセージ部分の情報を利用する. この部分は記事本文が書かれている. 文書分類においては, 形態素解析ツールなどにより, 単語あるいは文字の抽出を行ない, 得点をつけて特徴を調査し, 分類するのが一般的である. しかしながら, 対話型ニュースグループの場合, 記事本文中にかなりノイズが含まれているため, 形態素解析を行なっても, それほど精度が向上するとは思えない. 現在は処理すべき記事量が多いことから, 文字種成分の変化, つまり漢字とカタカナの連続部分を抽出し, その出現頻度といくつかの機能語を利用したスコアリングを行ない, スコアの高いタームを利用して関連度計算をする.

まず, このシステムでは抽出したターム中の文字列の頻度や表層的手がかりのみを用いて話題転換記事を決定している [7]. これは, 似ている話題が続いている場合は新出文字が少ないこと, 話題が変化する記事があれば, その記事の前後で出現する漢字の要素が異なるなどのヒューリスティクスを用いている.

類似度計算については, リファレンスツリー毎に出現ターム列にスコアを与え, スコアの上位 10 タームを利用する. ユーザが興味のある記事を選択した時点で, その記事を含むリファレンスツリーを類似計算の中心のツリー, ファミリーツリーとする. そして, ファミリーツリーの上位 10 タームを元に共起しているタームのスコアを利用したベクトル計算によって類似度を算出する [8].

2.3 知的ニュースリーダ -HISHO- の構成

HISHO システムアプリケーション版の構成は図 1のようになっている [9]. 大きく分けて, ニュースリーダ部と言語処理部にわかれ, 言語処理部においてはデータベース管理機能, 言語処理機能があり, その機能はいくつかのモジュールで構成されている. この部分において, 前章までに説明したデータ処理がなされる.

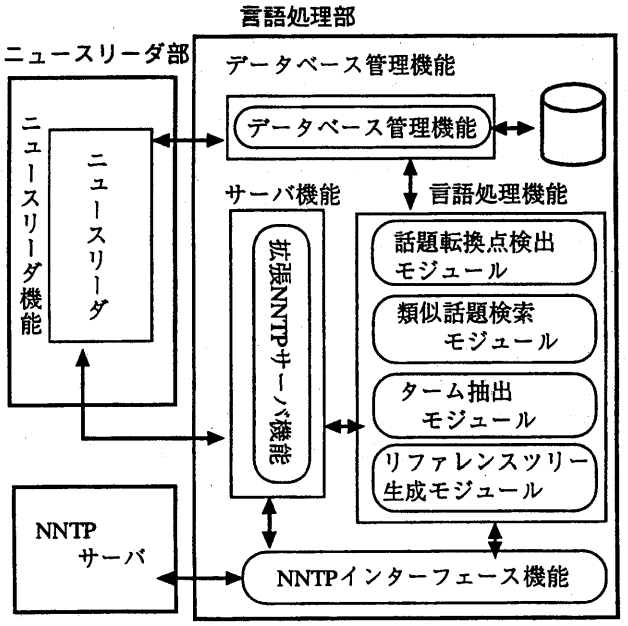

図 1: システム構成

一方, ニュースリーダ部においては, グラフィカルなユーザインターフェイスを提供し, リファレンスツリーや, 話題転換点の記事の表示, 類似リファレンスツリーの表示を行なう (図 2). アプレット版では, ブラウザを通して結果を表示していたが, 現在構築中の HISHO システムアプリケーション版ではブラウザの必要がない. また, 入力機能を充実させて, 次の話題の初めの記事に移動することなどが容易に行なえるようになっている. また, 表示ウィンドウの構成に改良を加えている.

FT No.	記事数	内容	関係する RT 数
950606-03	5	アトピー対策一般	11
950916-01	5	糖尿病	2
951213-03	3	いびき対策	2
951225-01	7	アトピーケアー	15
960130-07	2	風邪	3
960416-03	5	タバコと健康	3

表 1: ファミリーツリーの内容

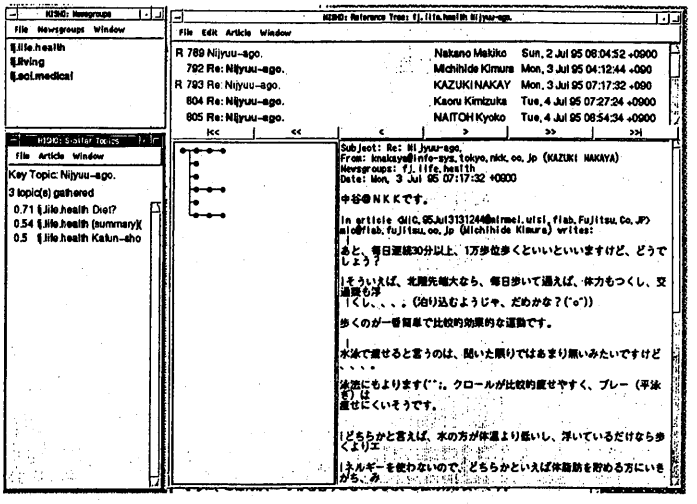

図 2: 知的ニュースリーダ -HISHO- アプリケーション版

3 知的ニュースリーダ -HISHO- の評価実験

現在, 我々は知的ニュースリーダ -HISHO- を JAVA 言語を利用して構築中である. 平成 9 年度に HISHO システムアプレット版は完成している. また, 今後の公開をめざし, HISHO システムアプリケーション版を構築中である. アプリケーション版では表示方法と類似リファレンスツリーの収集部分に改良を行なっている.

ここでは, 類似リファレンスツリーの収集方法の改良と評価実験の結果を報告する. なお, 評価実験に利用したニュースグループは fj.life.health, 記事は 1994 年 12 月から 1996 年 4 月までのものである. これらは, すべて北陸先端科学技術大学院大学 (JAIST) で立ちあげているアーカイブサーバ [10] から入手したものである.

3.1 収集方法とその評価結果

HISHO システムは, ターム抽出およびスコアリングには独自のツールを利用している [11]. このツールで, 漢字

FT No.	950606-03		950916-01		951213-03		951225-01		960130-07		960416-03		平均	
Methods	再現	適合	再現	適合	再現	適合	再現	適合	再現	適合	再現	適合	再現	適合
applet	10/11	10/21	2/2	2/19	2/2	2/20	11/15	11/19	3/3	3/16	3/3	3/13	—	—
	91	48	100	11	100	10	73	58	100	19	100	23	94	28
appli	10/11	10/18	1/2	1/2	1/2	1/1	15/15	15/17	2/3	2/8	2/3	2/2	—	—
-cation	91	56	50	50	50	100	100	88	67	25	67	100	71	70

表 2: 再現率と適合率

とカタカナからなる各タームに出現頻度と機能語によりスコアをつけている. これは, 各記事毎の得点であるので, アプレット版においては, このスコアの上位のタームをキーとして利用するが, 比較に使う際のスコアは tfidf による再計算を行ない, その得点を利用した. アプリケーション版では, ターム抽出の際に得たスコアをリファレンスツリー毎に記事数で正規化し, 比較に利用している.

次にテストデータによる収集実験の結果を報告する. 実験に利用したツリーは JAIST のデータから適当に抜粋した 34RT(200 記事) を利用している. その実験に利用したファミリーツリーは表 1のようになっている. このうち, アトピー関係のツリーでは 7 ツリー分重複している.

まず, すべてのツリーに 2 つの方法により得点を与える. その得点を利用し, ベクトル空間法を利用して類似度を計算する. ここでは, おのおののツリーから上位 10 タームを利用し, 同じタームが共起すると得点が高くなり, 似ていると判断される.

$$Sim(FTi, RTj) = \frac{FTi \times RTj}{|FTi| \times |RTj|}$$

この実験においては,2 文字以上の漢字文字列, カタカナ文字列と若干の機能語を利用し, 機能語の前にある 1 文字を抽出タームとしている. 類似度の得点が閾値を越えたものを正解とし, 人為的に選んだ結果と比較し, 再現率と適合率を算出した. 結果は表 2に示す.

$$\text{再現率} = \frac{\text{システム出力中の正解ツリー数}}{\text{正解ツリー数}}$$

$$\text{適合率} = \frac{\text{システム出力中の正解ツリー数}}{\text{システムが出力したツリー数}}$$

3.2 考察

テストデータにおいては, 再現率はアプレット版の方が高くなっている. しかし, 適合率がかなり低いため, 不要な結果を駆除するために表示方法を工夫する必要がある. また,tfidf のスコアリングは動的なデータに対して, どの段階で計算を行なうか, 全体量をどこに設定するかなど, スコアが大きく異なる要因をたくさん含んでいる. また, ニュースサーバ内の記事の変化に応じて, ツリー構造の変化にも対応しなくてはならない. その点, アプリケーション版においては, 各記事がサーバに届いた段階でタームのスコアを決定でき, ツリー構造の変化したツリーのスコアのみ再計算すれば良い. よって計算コスト的な面においてもかなり期待ができる. また, 適合率の向上により, 不要な結果が大幅に削減できる.

4 まとめ

知的ニュースリーダ -HISHO- システムの概要とシステム内の類似記事収集ツールの評価結果を報告した. これにより, ユーザのニーズに合わせた適切な議論の流れを提示することが可能となり, ユーザを適切に支援できる.

現在, アプリケーション版の構築を行なっている. その完成後, 表示部分の改良も含めて, ユーザインタフェースの評価を行なう予定である. また現在のシステムにはクロスポストの処理ツールが組み込まれていない. 今後はクロスポストおよび幅広いニュースグループに対応していく予定である. 最終的には平成 11 年度に一般に公開できるシステムにする予定である.

参考文献

[1] WIDE Project:"インターネット参加の手引" 共立出版,1995

[2] 五十嵐幸雄:"解説:情報検索" 日経エレクトロニクス No.705,1997.12.15

[3] Rennison, E.:"Galaxies of News: An Approch to Visualizing and Understanding Expansive News Landscapes" Proceedings of UIST94,1994

[4] 佐藤円他:"電子ニュースにおけるダイジェスト機構の実現" 情報処理学会第 49 回全国大会,1994

[5] 小作浩美他:"話題関連性に着目した知的ニュースリーダの提案" 平成 7 年電気関係学会関西支部連合大会,1995

[6] RFC 1036:"Standard for Interchange of USENET Messages"

[7] 内元清貴他:"対話型ネットニュースグループにおける話題転換点の推定" 言語処理学会第 3 回年次大会,1997

[8] 小作浩美他:"知的ニュースリーダにおける表層的話題関連性の抽出" 言語処理学会第 2 回年次大会,1996

[9] 井佐原均他:"討論型ネットニュースグループを対象とする知的ニュースリーダの開発" 情報処理学会,NL-119-3,1997

[10] http://mitsuko.jaist.ac.jp/fj/

[11] 宮本義男他:"キーワード抽出自動システム" 第 37 回システム制御情報学会研究発表講演会,1993