討論型記事群の一般用語出現傾向と知的ニュースリーダ-HISHO-への応用

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

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, かなりの量を維持している (図 1). そのため, 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].

それに伴い, いろいろな検索ツールが開発されている [2]. 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 [3] を行なってきた. このシステムは, 討論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.

HISHO システムでは, 内部計算において, 記事本文から抽出した, 漢字やカタカナの連続からなるタームを利用している. 記事量の多さから, 現在は形態素解析を行なっていない. また, 討論型ニュースグループにおいては, 話題が次々と変化していくため, 一般的に使われるタームが場合によって重要なキー, 話題の中心語となることもある. そのため, 一般用語辞書を利用してのキーワードの制限を行なわなかった. しかし, 類似話題記事の収集において, 記事の特徴語として一般用語が抽出されてしまい収集精度を下げていることも事実である.

そこで, 本稿では討論型ニュースグループにおける一般用語の出現傾向の調査結果を報告する. さらに, その出現傾向によって作成した一般用語リストを利用した,HISHO システムの類似話題検索評価実験の結果も報告する.

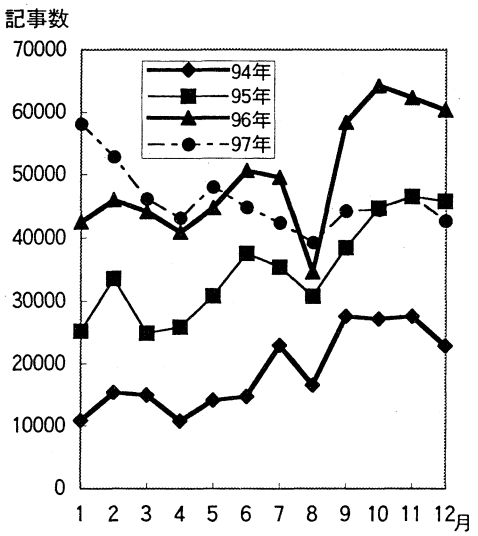

図 1: 投稿記事数の変化

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.

リファレンスツリー (RT) は記事の引用関係を利用して収集した記事群である. その構造は木構造をしている.

このシステムの主要な特徴は, ユーザの興味に沿って記事を動的に検索することにある. また動的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できる [4]. 次章では, 動作手順の 3 にある, ツリー (記事群) 同士の関連度計算について説明する.

表 1: 調査対象ニュースグループと記事数

年	ニュースグループ	記事数
95	fj.living	4796
97	fj.living	5520
95	fj.life.health	1596
97	fj.life.health	2826
95	fj.sci.medical	2006
97	fj.sci.medical	1595

3 関連度計算

関連度計算は 2 つのフェーズで行なわれる.RT 内の記事同士の関連度を調査する**話題転換点の決定**と RT 同士の関連度を調査する**類似話題の収集**である.

これらの計算においては, 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, 討論型ニュースグループの場合, 記事本文中に標準的な文でないものが含まれる上, 処理すべき記事量が多いため,HISHO システムでは形態素解析を行なっていない. その代用として, 文字種成分の変化, つまり漢字や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 [5], スコアの高いタームを利用して関連度計算をする.

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する [6]. 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, その記事の前後で頻出する文字が異なるなどのヒューリスティクスを用いている.

類似話題の収集では,RT 毎に出現ターム列にスコアを与え, スコアの上位 10 タームを利用し,RT 同士を比較する. 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 RT を類似計算の中心のツリー, ファミリーツリー (FT) とする. FT の上位 10 タームを元にタームのスコアを利用したベクトル計算によって, FT と他の RT との類似度を算出する [7]. この類似度計算の対象から一般用語を除くことにより精度の向上が期待されるが, 今までは一般用語辞書を利用していなかった. 討論型ニュースグループでは一般的なタームも場合によっては話題の中心語となる可能性があると考えていたためである. しかし, 類似度計算の際に障害になるタームが存在していることも明らかになってきた. それらは “私” や “投稿” などの一般用語と考えられるタームであり, 出現数がかなり多い. そこで, 討論型ネットニュースにおける高頻度タームを一般用語としてとらえ, その出現傾向について調査を行なった. 次章では, その調査結果を報告する.

4 討論型ネットワークニュースの一般用語出現傾向

この調査において fj.life.health, fj.living, fj.sci.medical の 1995 年 1 月から 1995 年 12 月までの記事と,1996 年 12 月から 1997 年 11 月の記事, および,1997 年 10 月一カ月間に 100 記事以上投稿のあった 90 の討論型ニュースグループの記事を利用している. 1995 年の記事は, 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 [8] から入手したもの, それ以外は当研究所で立ちあげているニュースサーバに届いた記事である. 各ニュースグループにおける記事数は表 1のようになる.

1997 年 10 月一カ月の間に 100 記事以上の投稿があった討論型ニュースグループは 90 ニュースグループあり,27389 記事であった. ニュースグループに依存しない高頻度タームを抽出する際, 各ニュースグループ毎にターム抽出を行なったが, その際はこの記事すべてを利用した. 実際, この記事はニュースサーバのスプールの問題上, クロスポストされた記事をコピーしているものがあるので, メッセージ ID を比較し, 重複するものがないようにした. その時の記事数は 25048 記事であった. 重複のない記事 25048 記事を総記事利用の際の調査対象とした.

なお,1997 年 10 月の時点で当研究所に存在した fj のニュースグループは 365 であり, 一カ月の総記事数は 42665 記事であった.

4.1 同じニュースグループ同士のターム比較

まず, 同じニュースグループで投稿された時期の違う記事に出てくる高頻度のタームについて比較を行なった. 利用した記事は表 1の記事である. 各ニュースグループ内の記事より漢字, カタカナからなる文字列を抽出し, その出現回数を投稿された年毎に集計した. その出現回数の上位 50 位までのタームを同じニュースグループの 95 年と 97 年の記事で比較した.

英語の文字列もタームとして切り出した場合,life.health においては 74%が一致しており,sci.medical では 58%, living では 50%が一致していた. 英語の文字列をタームとしない場合では, life.health で 70%, medical で 56%, living で 52%が一致していた.

同じニュースグループであれば, 時期にそれほど影響されない一般用語が存在することがわかる. また,life.health ではかなりニュースグループに特化した話題が投稿され

表 2: 一致しなかった高頻度ターム

年	グループ	ターム
95	fj.living	郵便局 銀行 アメリカ 値段
97	fj.living	携帯電話 障害者 電車 迷惑
95	fj.sci.medical	蔵器移植 脳死 判定 輸血
97	fj.sci.medical	肺癌 喫煙者 タバコ 根拠

表 3: 出現ターム比較

	総記事利用	グループ別	総記事−英語
95living	52%	50%	60%
97living	58%	54%	60%
95medical	46%	52%	50%
97medical	40%	46%	52%
95health	62%	58%	50%
97health	52%	60%	58%

るので, 一般用語的に利用されるタームの割合が大きいと考えられる. sci.medical や living においては, はばの広い話題が出されるため, 一般的に使われるタームが少ない.

一方,sci.medical と living において 95 年と 97 年とで一致しないタームを調査すると, それぞれ, その時期に特に話題として上がったものが見受けられる (表 2).

4.2 ニュースグループに依存しない一般用語

1997 年 10 月一カ月の間に 100 記事以上投稿のあった討論型ニュースグループの記事 (重複のないもの) をすべて調査し, 出現数の上位 50 のタームを抽出した (**総記事利用**). また, 重複した記事を利用し, 各ニュースグループ毎に出現タームを調査した. 続いて各ニュースグループの上位 50 のタームを比較していくつのニュースグループにそのタームが現れているか調査し, その上位 50 のタームを抽出した (**グループ別**). さらに, それらのタームリストと前章で利用した 3 つのニュースグループのリストをそれぞれ比較した.

一致したタームの割合を表 3に示す. 表 3において, **総記事利用**と**グループ別**は英語のタームも取り除かずに比較した時の一致した割合である. **総記事−英語**は総記事利用のタームリストと前章で利用した 3 つのニュースグループのリストから英語のタームを除いたもの同士を比較した時に一致したタームの割合である.

以上の結果より, 一つのニュースグループにおいて 1 年分の記事を利用して, 高出現頻度のタームを収集すれば, 半分以上がニュースグループに依存しない一般用語であると考えられる.

表 4: 頻度より抽出した一般用語

問題	方	必要	投稿	人	情報	気	理由	日本	結果
場合	実際	自分	私	記事	質問	現在	出来	意味	
意見	方法	使用	話	本当	事	普通	最近		

95 年,97 年の 3 つのニュースグループおよび総記事利用の英語を除いたタームリストを比較し, 半分以上のリストに現れているタームを収集すると表 4のようになる. これは, ニュースグループに依存しない一般用語であると考えられる.

以上のことから, ネットニュースにおいて高出現頻度のタームの上位 50 を調査すると, 高出現頻度のタームは大きく 3 種類に分けられることがわかる. 一般用語と同等に扱えるニュースグループに依存しない高出現頻度のターム (表 4),fj.life.health のようにあるニュースグループに特化して高頻度に出現するターム, 表 2のようにあるニュースグループにおいて話題の中心であるために高頻度出現しているタームである.

5 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. 既に 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 RT の収集部分に改良を行なっている.

ここでは, 類似 RT の収集の際に, 前章で抽出した高頻度タームを一般用語辞書として利用し, 評価した結果を報告する.

実験に利用したツリーは JAIST の fj.life.health のデータから適当に抜粋した 34RT(199 記事) を利用している. その実験に利用した検索キーとなる FT は表 5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.

一般用語辞書として,1995 年の fj.life.health の記事から抽出した高出現頻度上位 50 タームのリストと,1997 年の fj.life.health の記事から抽出した上位 50 タームのリスト, および前章で抽出したニュースグループに依存しない高出現頻度のタームのリスト (表 4) を利用した.

まず, テストセット内の各記事においてタームを抽出し, スコアをつける. 一般用語辞書を利用することで,RT の特徴キーとして利用できない高頻度タームが削除される. その結果, 記事の特徴キーはより記事にあったものが抽出

表 6: 再現率と適合率

FT No.	950606-03		950916-01		951213-03		951225-01		960130-07		960416-03		平均	
	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
95health	91	56	50	50	50	100	100	88	67	25	67	100	71	70
97health	82	50	50	50	50	100	100	88	67	25	67	100	69	69
independ	82	53	50	25	50	33	93	88	67	25	67	100	68	54

表 5: ファミリーツリーの内容

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

されている. 続いて記事のスコアを RT 毎に合算し, スコアの上位 10 タームをその RT の特徴キーとする. その特徴キーの得点を利用し, ベクトル空間法によって, 検索のキーとなる FT とすべての RT の類似度を計算する. ここでは, 各々のツリーから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.

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

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

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, 英単語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 6に示す.

5.1 考察

この再現率, 適合率を評価するために, E尺度による再現率, 適合率の重み付き平均を求め, 比較を行なった [9]. E尺度は次のような式で表される. ここでは, αは 0.5 とし, 適合率, 再現率は表 6の平均を利用した.

$$E = 1 - \frac{\text{適合率} \times \text{再現率}}{\alpha \times \text{適合率} + (1-\alpha) \times \text{再現率}}$$

E尺度の結果は,1995 年からのタームリストを利用した場合が 0.295, 1997 年からの場合が 0.310, グループに依存しないタームリストを利用した場合は 0.398 であった.

グループに依存しない一般用語のみを利用した場合は, そのニュースグループに特化した一般用語の影響が残ってしまい, 適合率が悪くなってしまう. ニュースグループに特有の一般用語の利用も必要である. また, 話題の中心であるため出現頻度が増えているタームについても抽出し処理する必要がある.

6 まとめ

討論型ネットワークニュースにおける一般用語の出現傾向について報告した. 高出現頻度のタームはニュースグループに依存しないものと依存するものとが存在し, ニュースグループに依存しない高出現頻度タームは一般用語と考えられることがわかった. さらにその高出現頻度のタームリストを利用し, 知的ニュースリーダ -HISHO- システム内の類似記事収集ツールの評価実験を行ない, 結果を述べた. 一般用語辞書の利用により収集結果を改善できることも示した.

しかし, 動的なニュース記事において, 一般用語辞書を固定したできるとは思えない. また, 期間や記事量を考慮した調査は行なっていない. より詳細な条件での調査が必要であろう. 今後は, 処理速度などへの考慮も行ない, より詳細な条件での調査結果も合わせて, 一般用語辞書をダイナミックに作成する方法について, 検討していく予定である.

本研究の一部は, 情報処理振興事業協会 「独創的情報技術育成事業」の一環として行われたものである.

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] 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

[4] 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

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

[6] 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

[7] 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

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

[9] "自然言語処理と検索技術講習会資料" 電子情報通信学会,1997