自然言語処理 117－4
（1997. 1. 20）

知的ニュースリーダが対象とする対話型ネットニュースの特徴

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- を開発している．その過程において，システムの取り扱うニュースグループの表層的特徴について調査し，その利用方法を検討した．

本稿では，その基本的な特徴について明らかにし，それを記事群の文脈抽出に利用するための若干の考察を加える．

Features of Discussion Type Network News Groups which was Handled by an Intelligent Network News Reader

Hiromi OZAKU　Kiyotaka UCHIMOTO　Hitoshi ISAHARA
Kansai Advanced Research Center,
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-. HISHO attempts to extract news articles which attract user interest. In the process of developping the system, we investigated several basic features of the network news groups.

In this paper, we report the features and consider ways to apply the features to the system.

1 はじめに

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. また阪神大震災直後の情報伝達においても, その有効性が評判となった. 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.

特にネット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する. そして, それらの中から必要なニュースグループを選択し, そのグループのすべての記事をチェックして必要なものを見つけるには相当な時間がかかってしまう. さらに, 出張等で数日ネットニュースにアクセスできなかった場合などは, 必要な情報を含む記事は記事の山に埋もれてしまうことになる.

また, 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も少なくなく, 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, 関係ある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.

そこで, 我々は, ネット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の提案 [1] を行なってきた. このシステムは, 記事の話題の流れを追い, ユーザの必要とする話題に関連する記事群をニュースグループを気にすることなく抽出することを目的として構築されている. そのシステム構築に際して, ネットニュースの表層的な特徴を調べ, 文脈抽出にどのように利用できそうか考察を行なった. 本稿では, その結果を報告する.

なお, 調査に利用したニュースグループは fj.life.health と fj.living, 記事は 1994 年 12 月から 1996 年 4 月までのもので, 記事数はそれぞれ 2346 記事, 7128 記事である. これらは, すべて北陸先端科学技術大学院大学で立ちあげているアーカイブサーバ [2] から入手したものである.

2 ネットニュースの特徴

最近のニュース記事の投稿量は非常に増えてきている.fj.news.lists に流れる news@nttiros.nslab.ntt.jp からの投稿記事数の集計記事を各月毎に集計すると, 投稿記事数は以下のように年々増加していることがわかる (図 1).

これにより, 最近は 1 日に 2000 件近い記事が投稿されていることがわかる. 1995 年のデータで, 1 日 1200 件, 約 3MB とあるので [3], 単純計算では, 1 日約 5MB もの記事が投稿されていることになる. この量の記事を毎日読むことは極めて困難であり, ネットニュースの記事を読むにあたり, 何らかのフィルタリング技術や検索技術が不可欠である. そして, その技術を開発するためには, その検索対象を良く知る必要がある. 次にその特徴の調査結果を報告する.

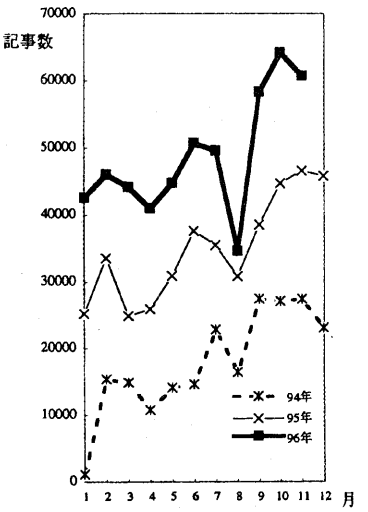

図 1: 投稿記事数の変化

2.1 ニュースグループ

一般にネットニュースには, 大きく分けて, 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ある [4]. fj のグループの場合, fj.news.lists 等に流れるニュースリストの一覧を調べてみると, 1996 年 10 月 28 日 (メッセージ ID < fjan6aa@cow.nara.sharp.co.jp >) の段階でグループ数は 327 個あり, そのうち, 287 個が対話型のニュースグループである. ほぼ 9 割が対話型である.

また,fj のニュースグループ数は年々増加の一途をたどっている (表 1). その分類は, わかり易く, 使い易いように考えられているようだが, ユーザには, そのグループの分類理由がうまく伝達されず, 逆にマルチポストや関係のないニュースグループへの投稿が増えているように思われる (ノイズの増加). また, ニュースグループ数が増えたことにより, 明確に分類しにくい場合も存在し, 記事内容により, 記

年月	グループ数	年月	グループ数
94.3.	184	95.7.	267
94.6.	201	95.9.	281
94.7.	220	95.10.	285
94.8.	221	95.11.	290
94.10.	222	96.1.	291
94.11.	224	96.2.	297
94.12.	226	96.3.	299
95.1.	230	96.6.	305
95.2.	233	96.7.	308
95.3.	243	96.8.	312
95.4.	244	96.9.	317
95.5.	250	96.10.	326
95.6.	253		

表 1: ニュースグループ数の変化

事の投稿されるグループが変わるケースも増えてきている. そのため, 興味のある記事を見つけてもその話題に関係する記事すべてを見つけるためには, いくつかのニュースグループをチェックしなければならなくなってきている (記事の分散).

最近の WWW の普及に伴い, アナウンス型のニュースグループの記事量の増加が停滞しているとの報告 [5] もあるため, これからますます対話型のニュースグループを対象とした検索技術やフィルタリング技術が必要になると思われる.

2.2 ニュース記事

ネットニュースの記事は, 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.

ここでは,Message-ID と References と Subject の各フィールドの特徴について報告する.

2.2.1 Message-ID と References フィールド

Message-ID は記事を投稿するにあたり, 必ず一意に決まるようになっている. そして, ある記事を参照したり, 何らかの回答やコメントを行なった場合, References に参照した記事の Message-ID が自動的に記述される. 故に, 記事同士の関係は Message-ID と References の関係を追うことである程度掴むことができるはずである.

記事同士の関係はツリー構造となる. 我々はこれをリファレンスツリーと呼んでいる. 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, たとえば,fj.life.health においては, 720 のツリーができるが, そのうちで 49 のツリーのトップ記事に, References フィールドが存在した. 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 (このような例は実際には 41 例存在した) 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 (8 例) である.

このような場合, 一般には, これらのツリーを関係付けることはできないが, 複数のツリーのトップ記事がニュースサーバに存在しない同じ記事を参照していた (同じ Message-ID を References に持つ) 場合には, それらを同じツリーにまとめることができる. この結果, トップ記事に References フィールドが存在する 49 のツリーは 42 のツリー (記事数は 139 記事) にまとめることができた. この結果, ツリーの総数は 713 となった.

同様に fj.living においては, トップ記事に References フィールドが存在する 121 のツリーを 105 のツリー (記事数は 515 記事) にまとめることができ, この結果, ツリーの総数は 1049 となった. なお, この中で, 話題が他のニュースグループから移動してきたものは 84 例, 取り扱った期間以前から継続している話題のものが 24 例, ニュースサーバに記事が届いていないと思われる場合が 13 例であった.

次にリファレンスツリーに含まれる記事数であるが, それぞれ (表 2,3) の通りである. ツリーを構成する平均記事数はそれぞれ 10.5 記事と 6.7 記事であった.

2.2.2 Subject フィールド

Subject フィールドには, 記事のタイトルが記述してある. ある記事を参照して記事を投稿すると,

記事数	ツリー数	記事数	ツリー数
1	380	15	1
2	99	16	2
3	56	17	1
4	36	19	2
5	37	20	2
6	15	22	2
7	12	24	2
8	14	27	1
9	11	28	1
10	9	30	1
11	5	38	1
12	9	51	1
13	4	61	1
14	8		

表 2: fj.life.health のリファレンスツリーサイズ

ユーザが何らかの手を加えない限り Subject には参照記事のタイトルに Re:がついた形で保持される. つまり, 関係のある記事同士はほぼ同じタイトルを持つことになる.

Subject の特徴としては, fj.life.health においては季節に依存するタイトルは周期的に出易い (タイトル例 :風邪, インフルエンザ, 花粉症など). また, このグループの場合, 健康に関するタイトルは比較的繰り返し出てくる (癌の治療, 視力回復, アトピーなど).

fj.living においては, さすがに生活全般の話題を扱うグループであるためか, 引っ越しの話題や生活習慣に関する話題 (引出物, 御歳暮など), また常日頃困っていること, 苦情に関するもの (間違い電話, 訪問販売など) も出易い.

fj.life.health の記事において, 話題提起 (Re:あるいは re:のついていない) 記事から, Re: あるいは re: がついただけの違いの記事を抽出し,Subject の持続期間と記事数を調査した. 対象記事数は 1913 記事, タイトル数は 708 タイトルであった. さらに, その 708 タイトルのうち話題提起記事のタイトルを含むもの, 例えば atopi と [Summary]atopi のようなものを同じタイトルとして扱う. するとタイトルの数は 315 となる.

同じ Subject が続く記事の投稿始めの日から投

記事数	ツリー数	記事数	ツリー数
1	437	31	2
2	132	32	2
3	93	33	2
4	67	34	3
5	38	35	1
6	48	38	1
7	23	39	1
8	27	40	1
9	14	41	2
10	23	43	1
11	9	46	1
12	19	47	1
13	14	53	2
14	7	54	3
15	9	55	1
16	5	56	1
17	6	59	1
18	5	61	2
19	4	73	1
20	2	90	1
21	4	93	1
22	3	94	1
23	2	97	1
24	3	100	1
25	3	111	1
26	1	150	1
27	4	154	1
28	5	238	1
29	1	407	1
30	2		

表 3: fj.living のリファレンスツリーサイズ

稿終りの日までをそのタイトルの持続期間とし, 持続期間と記事数及びタイトル数の関係を調査した (表 4). 一番長い期間は 96 日であった.

3 文脈抽出のための考察

大規模な記事において, 基本的な記事の関係を掴むには, 投稿者が明示的に行なった引用関係を示す, リファレンスツリーを利用するのが効率的である. ここでは, リファレンスツリーを中心に, 知的ニュースリーダ -HISHO- を構築するために行なった考察を述べる.

3.1 Message-ID と References フィールド

当初, 当研究所のニュースサーバにおいて, 記事を収集することを試みた. ところが, 当研究所のサーバは立上げ時期であった事もあり, かなりの記事がサーバに届いていなかった. 当初は, リファレンスツリーを利用して記事の関係を大まかに掴む事すら難しいと思われた.

しかし, ニュースサーバが安定するにつれ, 何らかのトラブルにより届かない記事量は激減してきている. これは, 単にニュースサーバが安定してきただけの理由ではなく, 最近のネットワーク運営管理が専門の会社等によって行われる事が増えて来たこともあり, データの安定供給が確実に行われるようになったためとも考えられる. これらのことより, ニュースサーバのネットワーク上の位置や転送経路にもよるが, 何らかのトラブルによりサーバに届かない記事はかなり少ないと考えられる. また, アーカイブサーバを利用することにより, 失った記事の収集もできると考えられる. よって, リファレンスツリーをうまく活用する事により, 大まかな分類が可能であると思われる.

リファレンスツリーの構造として特徴的なのは, 今そのニュースグループで中心的な話題の記事には, より多くのユーザがコメントを与えるということである. つまり, リファレンスツリーの構造上,1 つの記事より多数の枝に分岐する. 大勢のユーザが議論に参加していれば枝分かれが多くなる分, 新しい情報が提示されるため, 話題転換も起こりやすい. また, 大幅な話題転換が行われる場合, 投稿者は Subject を変えて意図的に新しい話題提起記事を投稿する. つまり, リファレンスツリーも新しくなる. 同じ話題でもまとめ記事のようにしばらく後に投稿されるような場合, リファレンスツリーが異なりやすい.

さらに, リファレンスツリーでつながっている記事は, 引用した記事の部分コピーをする事が多いため, 話題のキーとなる単語は複数現れ易い [7]. また, まとめの記事には一連の話題を総括するため, 話題のキーとなる単語が現れ易い.

このような構造の変化と特徴をうまく利用する事により, ユーザーの必要とする話題がどのリファレンスツリーの枝に含まれるか, 詳細にチェックすべき部分を把握できるものと思われる.

我々は, これらを利用し, リファレンスツリー毎にいろいろな抽出方法でリファレンスツリーの構造とキーワードなどの特徴を取り出し, 類似点や相違点を把握するための比較実験を行っている.

3.2 Subject フィールド

表 4 を累積度数計算すると, 表 5 のようになる. これを見ると, 30 日で全体の 90 % を占めている. さらに, 40 日以内に 95 % の記事が集まっている. よって,Subject から記事間の関係をとる場合, キーとなる記事から前後 30 日か,40 日調べれば良いと判断される.

リファレンスツリーとの関係を見ると, リファレンスツリーの数より (713) も, Subject で分類した方 (636) が, 約 1 割ほど少ないグループに分類できる. これらを考慮すると,Subject でリファレンスツリー同士の関係を見付けることができる可能性がある. 実際,Subject が全く異なるのに, リファレンスツリーが同じになる例 (6 例) より, リファレンスツリーが異なるが Subject は同じである例 (37 例, 37 タイトル 84 リファレスツリー) の方が多い.

これは, 話題提起記事を投稿した人がいくつかの回答やコメントをもらい, その後しばらくの調査期間を経て, まとめの記事を投稿したような時に起こりやすい. また, 大きなリファレンスツリーの始めの記事が expire された時にも起こる.

リファレンスツリーの同士関係を低コストでとるのであれば,Subject を利用することも可能である. つまり, 話題提起の記事と関係のある記事は Subject が同じである可能性が高い. しかし, ユーザの記述の揺れなどより, 違う Subject になったり, MIME

保存期間	記事数									
	2	3	4	5	6	7	8	9	10	11以上
1	22	5	1	1	0	0	1	0	0	
2	26	16	2	3	1	1	0	0	0	
3	8	7	5	3	0	0	2	1	0	
4	7	4	1	0	0	0	0	1	1	11-1
5	3	4	2	3	4	1	0	0	1	
6	2	4	5	1	1	2	0	0	1	12-1
7	5	4	0	3	3	2	0	0	1	11-1,14-1
8	4	2	1	3	2	1	1	1	0	11-1,14-2
9	1	0	1	0	1	0	2	0	0	12-1,13-1,20-1
10	0	1	2	4	0	0	1	2	1	14-1,25-1
11	2	1	2	0	1	0	0	0	0	16-1,31-1,12-1
12	4	3	2	3	1	0	1	0	0	11-1
13	0	1	0	1	0	0	0	1	2	13-1,15-1,28-1
14	1	1	0	1	1	1	0	1	2	12-1,13-1
15	0	1	0	1	0	0	0	0	0	18-1
16	0	0	0	0	0	0	0	0	0	12-1,14-2,29-1
17	0	0	1	0	2	0	0	0	1	24-1,27-1
18	0	1	0	1	0	1	1	0	0	11-1
19	0	2	0	0	0	0	0	0	0	
20	0	0	0	3	0	0	0	0	0	
21	0	0	1	0	0	0	1	0	1	12-1
22	0	0	0	1	0	0	0	0	0	21-1
23	0	0	0	0	0	0	1	0	0	
24	1	0	0	0	0	0	0	0	0	11-1,17-1
25	0	1	0	0	0	0	0	0	0	12-1
26	0	0	0	0	0	0	0	0	0	24-1
27	0	0	0	1	0	0	1	0	0	24-1
28	0	0	0	0	0	0	0	0	0	15-1,51-1
30	0	0	1	0	0	0	0	1	0	14-1
31	0	0	0	1	0	1	0	0	0	14-1
33	0	0	1	1	0	0	0	0	0	16-1
37	0	0	0	1	0	0	0	0	0	
40	0	0	0	0	0	0	0	0	0	20-1
46	0	0	1	0	0	0	0	0	0	
53	0	0	0	0	0	0	0	1	0	
56	0	0	0	0	0	0	0	0	0	62-1
60	0	1	0	0	0	0	0	0	0	
79	0	0	0	0	0	0	0	0	0	19-1
98	0	0	0	0	1	0	0	0	0	

表 4: Subject の維持期間と記事数及びタイトル数

期間 (日)	記事数	累積記事数	比率 (%)
1	76	76	4.0
2	136	212	11.0
3	97	309	16.2
4	60	369	19.3
5	82	451	23.6
6	83	534	27.9
7	104	638	33.4
8	108	746	39.0
9	73	819	42.8
10	106	925	48.4
11	80	1005	52.5
12	65	1070	55.9
13	93	1163	60.8
14	77	1240	64.8
15	26	1266	66.2
16	69	1335	69.8
17	77	1412	73.8
18	34	1446	75.6
19	6	1452	75.9
20	15	1467	76.9
21	34	1501	78.5
22	26	1527	79.8
23	8	1535	80.2
24	30	1565	81.8
25	15	1580	82.6
26	24	1604	83.8
27	37	1641	85.8
28	66	1707	89.2
30	27	1734	90.6
31	26	1760	92.0
33	25	1785	93.3
37	5	1790	93.6
40	20	1810	94.6
46	4	1814	94.8
53	9	1823	95.2
56	62	1885	98.5
60	3	1888	98.7
79	19	1907	99.7
98	6	1913	100

表 5: Subject 持続期間と記事数

コード Subject による問題 (文字化けや単なるコードの羅列化) など, 低コストの比較だけでは関係を取り出せない部分も多々ある. また, 全く違う内容の記事を投稿しながら, たまたま同じタイトルになる場合も存在し, 現在のところ必ずしも明確な利用方法が提案できるとは思えない.

4 まとめ

対話型ニュースグループ,fj.life.health と fj.living の基本的な特徴についてまとめた. また, ネットニュースの効率的な利用のための考察も行った.

今後はより専門的なニュースグループについても検討を行いたいと思っている. さらに, これを基礎に知的ニュースリーダ -HISHO- の開発を行っていく予定である (図 2). また, それに伴い, システムの評価も必要になってくると思われる. 我々は評価用セットの作成も行なっている. これは, fj.life.health の 2346 記事中からいくつかキーになる記事を選択し, それを基に人手で類似しているものを抽出したものである. 今後は, その評価用セットを利用したシステムの評価についても行なう予定である.

参考文献

[1] 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

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

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

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

[5] 佐藤円他:"ネットニュースとダイジェスト自動生成" コンピュータソフトウェア, Vol.13, No.5,1996

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

[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

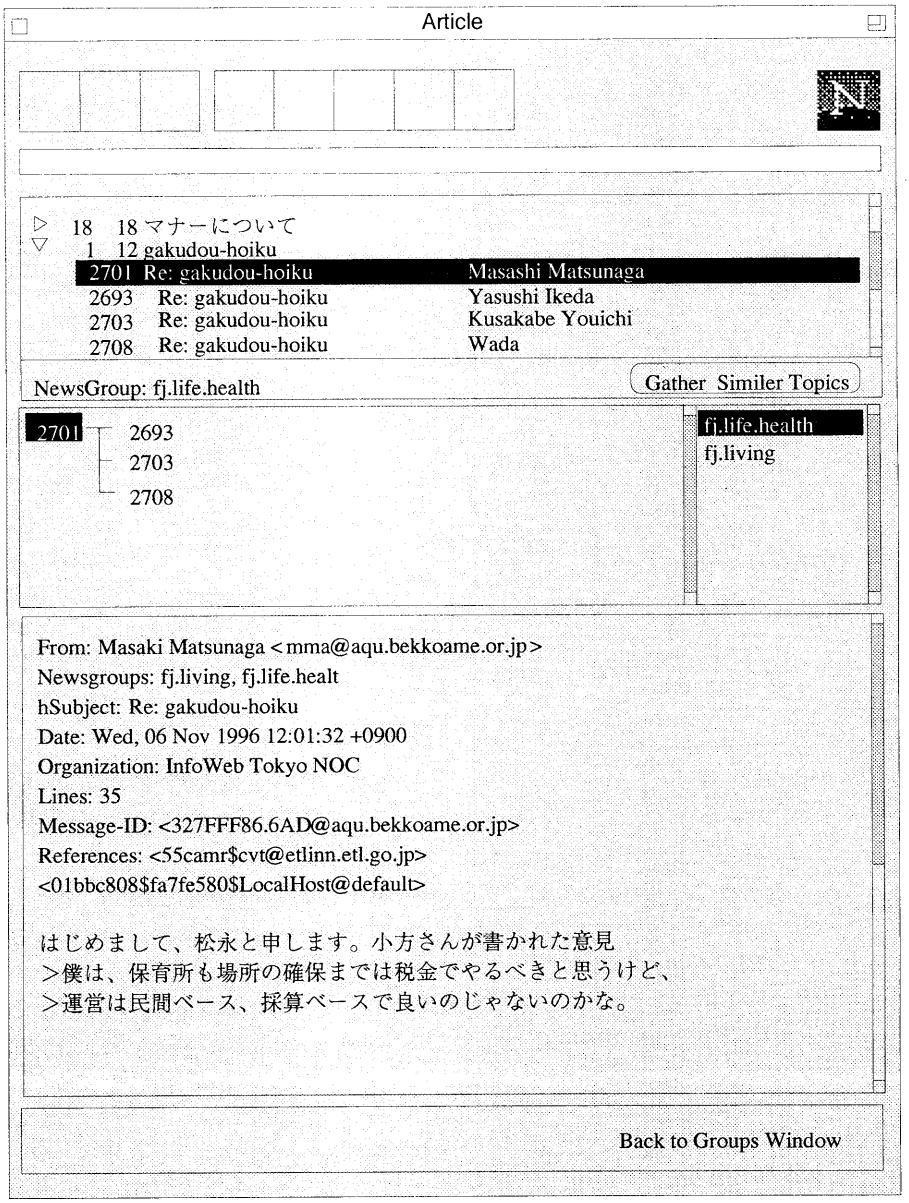

図 2: 知的ニュースリーダ -HISHO-