知的ニュースリーダにおける 表層的話題関連性の抽出

小作 浩美　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,isahara}@crl.go.jp

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 のグループの記事だけでも 1 日に約 1200 件, 約 3MB もの記事が流れており [1], それらをすべてチェックし, 必要な記事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 (以下 NG) 間を移動する話題も少なくなく, 適切な NG にアクセスして必要な情報を抽出するのはかなり困難となっている.

ネットニュースの情報をより効率よく利用するため, ダイジェスト自動生成の実現 [2] や知的ニュースリーダの提案 [3] もされ始めている. しかし, NG によっては, ダイジェストに向かないものもあり, すべての NG について取り扱える訳ではない.

ネットニュースには, 大きく分けて, 新聞記事型の NG と, 討論型の NG があり [4], 特に討論型のニュースにおいては, ユーザのニーズを考えた場合, ダイジェストで満足する場合は必ずしも多くはなく, 興味を持った内容については, 関連する記事そのものを読みたいであろうと考えられる.

そこで, 我々は討論型 NG を対象として, 話題関連性に着目した知的ニュースリーダの構築を行っている [5].

取り扱うべきニュース記事の量が膨大であることから, まず, 知的ニュースリーダの構築の第一段階として, 討論型の NG について表層的な特徴を調査し, それを利用して関連が少しでもありそうな記事を抽出し, そこからさらに高コストの処理によって, 必要な記事を抽出するということがどの程度可能であるか検証を行なった. 本稿では, その調査結果について報告する.

2 ネットニュース利用における問題点

人間が, ネットニュースに流れる情報を効率的に利用するには大きく分けて 2 つの方法が考えられる.

1. ネットニュース全体において情報検索をする.
2. NG にわけて情報検索をする.

ネットワークを流れる情報を十分にを利用しようとすると 1 が好ましいが, ネットニュースの情報量は膨大であり, その一つ一つに対して, 高コストの自然言語処理を行なうというのは現実的ではない.

また, ある NG を 1 つ選択し, その中の話題について調査するとしても, 個々の記事はクロスポストされている場合があり, いくつかの NG にまたがって話題が進展し, 必要な情報は別の NG に存在していることも少なくない. また, 面白い話題の存在する NG は往々にして投稿数が多く, 読み切るだけでかなりの時間を費やしてしまう. さらに, 話の流れによっては, 関係ない話題へと進んでいくこともあり, 全てを読んでも必要な情報があるとは限らない.

そこで, ダイジェスト自動生成システム等が構築されてきている訳であるが, ダイジェストは, ある程度決まった記述方式を取る新聞記事型 NG では有効であるが, 自然対話に近い記述方式を取る討論型 NG では, 内容をより深く解析し理解しないと一概に要約することはできない. また, そもそも, 要約することが意味をなさないことも多い.

また, 一般的な検索システムでは, ユーザがどのような記事を検索したいかをシステムに指示する手法としては, 簡単なキーワード入力によるものや質疑応答を行ないながら, 指示内容を深めていくものがある.

しかしながら, いくら全文検索が素早くできるとしても, どのような話題があるかわからない状態でキーワードを決めることは困難であり, また, コンピュータに慣れないユーザに対してはかなりの負担となる.

3 システムの概要

前章の問題を踏まえ, 我々は討論型 NG に対し, キーワードを入力することなく, 興味のある記事を検索できる知的ニュースリーダの構築を行っている. システムのイメージとしては, しばらくネットニュースにアクセスしていなかったユーザが, 久しぶりに記事を読み, その中から興味のある記事を発見した場合, その記事を選択するだけで, 関連のある記事群を検索し, 提示してくれるようなシステムである.

3.1 システムの動作例

システムの利用法は, 以下のようになると思われる.

1. ユーザが記事群の中に興味のある記事を見つける.
2. システムはユーザが興味を持った記事の意味的特徴を記事中の利用単語と, 概念辞書を用いて得る.
3. Subject,References 情報を用いて, 記事の関連ネットワーク (次節参照) を成長させていく.
4. 関連ネットワーク中の枝分かれについて, 記事に関するヒューリスティックスと, 元記事の意味的特徴を利用して, 関連の深い記事とそうでないものとに分類する. 関連の深い記事の関連ネットワークを成長させていく.
5. 元記事の意味的特徴をキーとして,References で継っていない関連記事をも抽出する.
6. 5. で得られた記事についても 3. 4. を行なう.

本稿で述べる手法の目的は, この 5 の過程において対象とする記事を予め減少させておくことにある.

3.2 関連ネットワーク

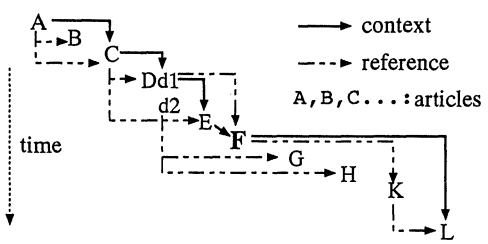

図 1: 記事と関連ネットワーク

図 1において, ユーザが注目した記事が F であった場合, その記事から Subject,References のリンクで到達できる記事を網羅する. 図 1中で, D の記事は 2 つの異なった話題について引用 (リファレンス) されている. このような枝分かれにおいては, F の内容と直接の関係のないもの (G, H) については, その先の検索は行なわない.

本システムにおいて, ニュース記事は最終的には次のように表示される.

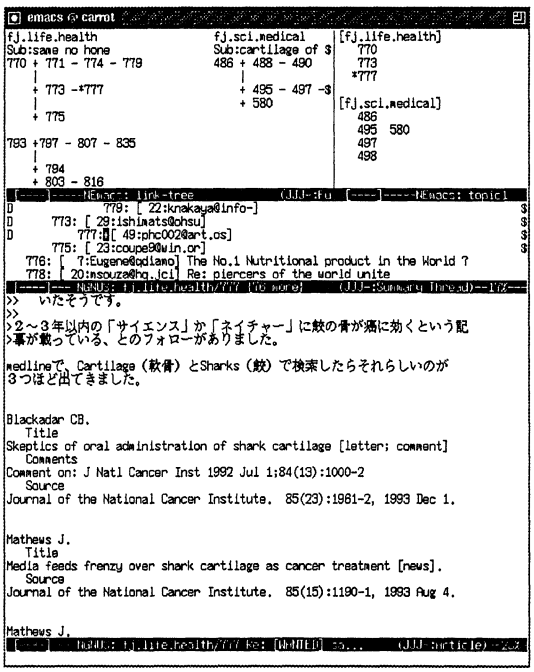

図 2: 動作例

4 表層的な特徴とその抽出

本稿では, 検索範囲をすばやく決定するための処理において, 選択された記事の表層表現からどの程度情報が抽出でき, また, それを利用して話題関連性がどの程度抽出できるか調査を行った結果について述べる.

ネットニュースの記事は, ヘッダとメッセージの 2 つから構成されている [6]. ヘッダはいくつのフィールドから構成され, 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 のフィールドがある. さらに, その記事のタイトルを記述する Subject のフィールドがある. これらのフィールドを利用すれば, 記事の関係がある程度掴めると考えられる. そこで, まず, このヘッダのフィールドを利用して関係のある記事を選別することを試みた.

記事数	fj.life.health	fj.living
1	191	309
2	98	60
3	65	35
4	45	20
5	30	18
6	31	14
7	12	12
8	17	7
9	10	8
10	16	5

表 1: リファレンスツリーの構成記事数とツリー数

通信総合研究所関西先端研究センターのニュースサーバマシンより, 2つの討論型 NG, fj.life.health と fj.living それぞれ 1 年分の記事を取り出し, References でどの程度記事を選別できるか調査を行った.

4.1 References による分類

既に述べたように NG には大きく分けて 2 つのタイプが存在する. 討論型 NG では, 記述方式が自然対話に近く, 対話の流れを示すために, 関係のある記事の関係ある部分を抜粋し, 引用するケースが多い.

そのため, 関係のある記事は引用部分を利用することで抽出することが可能である場合が多いと考えられる. 引用した記事については, References に Message-ID が記述されるので, References と Message-ID を利用して関係を抽出してみた.

fj.life.health においては, 1 年間に 1330 記事が投稿され, 引用関係にある記事をまとめることにより, 396 の記事群に分類出来た. この記事群は, 引用構造に沿った木構造をなす. この木構造をリファレンスツリーと呼ぶことにする.

fj.living においては, 4248 記事が投稿され, 724 のリファレンスツリーが生成出来た.

続いて, 各リファレンスツリーの構成記事数を調査した. (表 1に一部を示す) 構成記事数の一番多いものは, fj.life.health において 61 , fj.living では, 154 であった.

表 1でもわかるように, そのリファレンスツリーのうち, 他の記事と一切 References から関係が見つけられず, 単独記事でツリーを構成するものは, fj.life.health では, 191 ある. 個々のリファレンスツリーの内容を検討することにより, そのうち 54 は他の記事と何らかの関係があることがわかった. fj.living では, 309 が記事 1 つから構成され, そのうち 129 は, 他の記事と何らかの関係が見つけられた. また, 複数の記事で構成されるリファレンスツリーのトップの記事でも他のリファレンスツリーと何らかの関係のあるものも見受けられた.

Reference フィールドを利用しても関係を見つけられない理由として以下のことがあげられる.

1. ユーザが故意に References を削除したり, 記述しなかったため
2. ニュースサーバにより記事をエクスパイアされたため元記事が転送されなかったため
3. まとめ記事は, ある期間をおいて投稿され, 元記事をわざわざ引用しないため
4. References フィールドが長くなりすぎたため, 転送中に記述が壊れたため

以上のように, References を利用することによって, 関係ある記事をある程度選別することは可能であるが, 完全に分類できるわけではないことがわかった.

4.2 Subject による分類

別のリファレンスツリーとして分類されたものを関係付ける方法として, ヘッダの Subject フィールドを利用してみた. もし, 元々同じ記事から話が始まっているのであれば, 同じタイトル (Subject) である場合が多いと考えられる. そこで, リファレンスツリーの Subject 同士を比較することにより, リファレンスツリー相互の関係を抽出することがどの程度可能か調査を行った.

fj.life.health の 396 リファレンスツリーに使われている Subject の種類は 362 ある. fj.living においては 632 種類であった. 従って, Subject フィールドを利用すれば, 1 割ほど関係が抽出できることがわかった.

しかし, ある程度の期間, 同じ Subject が存在している場合, 内容的にその Subject に沿った話題が常に記述されているとは考えにくく, 不要な記事も含まれてくる. また, Subject もユーザによって記述方法が異なり, Subject から内容を全く連想できない場合 (“おためしあーれ” “教えて下さい” 等) や, 同じ話題でありながら全く違う Subject のもの (“ダイエット” “二重顎” 等) も存在し, Subject フィールドの利用による選別はかなり難しいことがわかった.

4.3 出現文字列による分類

同じ Subject であっても, 同じ話題の情報を記述しているわけでないとなると, メッセージの “内容” (話題) をある程度調べる必要がある. 異なる NG にも, 必要とされる記事がある可能性があることから, 記事の探索範囲は (概ね) 全 NG となり, 記事量が膨大となることから, 個々の記事を詳細に解析するのは現実的でない. そこで, メッセージ部分から, 漢字またはカタカナによる連続文字列の抽出を行い, その出現文字列になんらかの特徴がないか調査を行った.

まず, 常用出現文字列の調査を行った. これは, 常に出てくる文字を話題の中心語として処理してしまうのをさけるために行った. 各 NG において, 1000 記事を 100 記事づつの記事群に分けて出現文字列の調査を行い, 各記事群の上位 50 の文字列を取り出し, どの記事群にもほぼ現れている文字列を常用出現文字列とした.

fj.life.health においては, “思” “私” “場合” など約 20 文字列が, fj.living についても同様に約 20 文字列が常用出現文字列として抽出された.

5 表層的な特徴による関連性の抽出

次に, 我々は References 及び Subject のフィールドを利用して選択された記事群のメッセージ部分の出現文字列を抽出し, いくつかのリファレンスツリーにおいて比較を行った.

Subject が同じリファレンスツリーと, Subject は異なるが内容が似ていると思われるリファレンスツリーをそれぞれ構成する記事数に応じていくつか抜き出し, 出現文字列を調査し, それぞれ比較してみた.

出現文字列の調査においては, 漢字とカタカナであれば, 常用出現文字列以外を全て抜き出した. そして, それぞれの文字列の出現数を総出現文字列数で割った出現率を計算した. さらに, 出現率低いのものは出現文字列リストから削除した.

続いて, リファレンスツリー同士を比較し, 同じ出現文字列があれば, それぞれの出現数と出現率を掛け合わせ, 全てを足して関係量を計算した. その際, その総数が 100 以上であれば, 関係が深いもの, 50 未満のものは関係がないものとして分類し, 実際の記事との比較調査を行った.

それにより, 同じ Subject であっても, 違う Subject であっても, そのリファレンスツリーとの間の関係の有無が, ある程度, 判断できることがわかった.

6 まとめと今後の課題

本稿では, 記事の流れを判定するために, ネットワーク上の全記事群から, 簡易な手法で, ある程度のスクリーニングを行ない, その結果をより詳細な判定処理に利用しようという試みについて述べた.

本稿に示す様に, 表層的な特徴によって, リファレンスツリーを構成する記事の数に関係なくリファレンスツリーの関係を抽出出来ることがわかった. この手法により, ユーザの選択した記事に関連のありそうな記事群を取り出し, その記事群中の記事を減らす方向で, さらに話題の流れを調べれば良いものと考えられる.

今後の課題としては, 出現文字列のより詳しい調査を行い, 文字列の共起関係や統計的利用による記事の絞り込みの可能性を探る予定である. また, 抽出された記事群をより詳しく調査することにより, ユーザの興味に即した記事や, 話題の流れを把握する必要がある. よって, 抽出されたリファレンスツリー中の記事のメッセージ部分から引用部分を切り出し, 引用部分の特徴を利用して関係を詳しく調べる予定である.

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] 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” 情処第 49 回後期全国大会,1994

[3] 岩爪道昭他:“電子掲示版における記事の自動分類と議論の可視化ー知的ニュースリーダの提案” 人工知能学会全国大会,1994

[4] Rennison, E.:“Galax¡ies of News: An Approch to Visualizing and Understanding Expansive News Landscapes” Proceedings of UIST94,1994

[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”