利用者の興味を反映できる複数文書要約

村田 一郎 †　　**森 辰則** ‡
† 横浜国立大学 大学院 環境情報学府　　‡ 横浜国立大学 大学院 環境情報研究院
E-mail: {ichiro,mori}@forest.eis.ynu.ac.jp

1 はじめに

大量の情報の中から必要なものを効率良く探し出すために情報アクセス技術の研究が広く行なわれている．その 1 つである自動要約は，文書から不要な部分を削ることで利用者が読む量を減らし，文書情報を効率良く扱えるようにする技術であるが，同じ文書を要約するにしても利用者がどのような情報を求めているかによって提示すべき要約は変わってくる．さらに自動要約を情報アクセス技術として用いる場合，利用者は詳細な情報要求を持たず，興味が広い範囲にわたっていることが多いと考えられる．このような場合は利用者の興味を特定することが難しく，1 度の要約処理だけで十分な情報が提示できるとは考えにくい．しかし，何度も要約処理前の段階に戻り，条件を変えて要約し直すことは利用者にとって負担が大きく，またそれだけでは利用者の興味を反映することに限界がある．

本論文では利用者との対話的なインターフェースを通して利用者の興味を要約処理に反映し，提示する手法を検討する．特にそのインターフェースとして，提示した要約文章に利用者が直接操作を加えることで，直観的かつ簡単に情報要求のフィードバックを行なえる手法を提案する．

2 関連研究

2.1 利用者の興味に焦点を当てた要約手法

利用者の興味を考慮する要約手法としては，Tombros ら [1] の提案する Query-biased Summarization がある．これは文書検索結果に対する要約を行なう場合に，利用者が文書検索に用いたキーワードの重要度を高くして重要文抽出を行なうものである．また，利用者が質問文を与えた場合にそれを考慮した要約を提示する研究もあり，平尾ら [2] は質問が問うている事物の種類の情報を用いる手法を提案し，我々も質問応答エンジンの出力を用いる手法 [3] を提案している．

2.2 対話型要約手法

長尾ら [4] は GDA に基づく要約手法を提案し，対話的に要約を行なえる要約ブラウザを紹介している．これは要約に用いた原文書の中から利用者が単語や文を選択すると，それに関連した内容を重視して要約を表示し直すというものである．また，Leuski ら [5] は iNeATS という対話型の複数文書要約システムを提案している．このシステムでは利用者が要約の長さやトピックを指定することができ，出力した要約文について重要度の可視化などを行なえる．

2.3 Scatter/Gather 法

Scatter/Gather 法 [6] は，大きな文書集合の中から利用者にとって興味のある集合を絞り込む際に用いられる．これは文書集合をクラスタリングしてそれぞれのクラスタの代表語を利用者に提示する過程 (Scatter) と，利用者がそれらの中から興味のあるクラスタを選択して新たな文書集合とする (Gather) 過程からなり，それぞれの過程を繰り返すことで，利用者に必要な文書集合を効率よく得ようとする手法である．

2.4 本研究の位置付け

利用者の興味に焦点を当てた要約に関する先行研究では，いずれも要約処理前に利用者の興味を調べ，それに対する要約を提示した時点で処理は完結しており，提示した要約は最終出力という位置付けになっている．これに対し我々は，Scatter/Gather 法を要約提示の観点から捉え直す事により，提示した要約文章そのものに対し利用者が操作を行ない，それによって利用者の興味を反映した新たな要約を提示する手法を提案する．この利用者の操作とそれに対する要約提示という手順を繰り返すことで利用者の情報要求に対する適切な要約を提示できると考えられる．次章でこの提案手法について詳しく述べる．

3 Scatter/Gather に基づく利用者との対話による要約手法

3.1 基本的な考え方

利用者の質問に対してその答を提示する質問応答は，利用者から情報要求が詳細に与えられる場合に有効な情報アクセス技術である．これに対してある程度の長さの文章を提示する自動要約は，利用者の情報要求が定まっていないか，あるいは広い範囲にわたって興味がある場合に有効な情報アクセス技術となり得る．この場合，利用者に提示する要約としては初めに文書集合全体を概観するものを見せ，利用者の要求に応じて次第にトピックを狭めた要約を順に提示していくという方法が望ましいと考えられる．

ここで我々は，Scatter/Gather 法を要約文章提示の観点から整理し直すことにより，上記の手順を実現する一手法を提案する．特に，この手法では利用者の操作を要約文章に対する操作に集約できるという特徴がある．具体的には，Scatter 過程を，文書集合全体を概観する要約文章を生成し，利用者に提示することと捉え直し，Gather 過程を，提示された要約文章のうち，情報要求に適合する文章部分に利用者がマウス操作等により自由に印を付与することと考える．システムは印が付与された文章部分に関連する文書 (あるいは文書部分) を集め，これらを改めて要約対象文書として，複数文書要約を行うことを繰り返す．図 1に上記の処理の概要を示す．

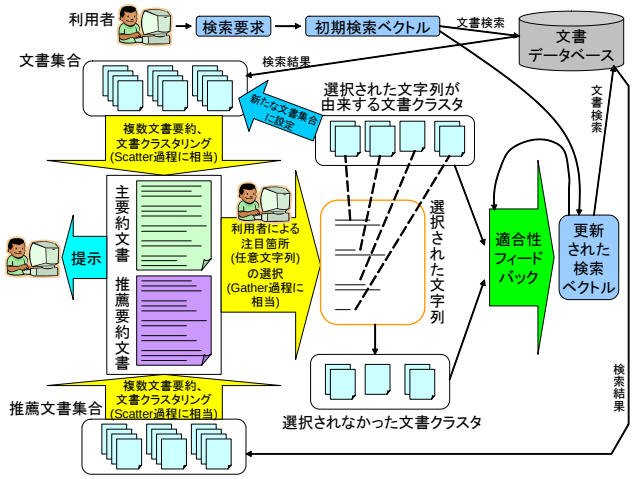

図 1: 対話型複数文書要約の流れ

Scatter/Gather 法と自動要約手法を組み合わせる従来手法においては，情報アクセスの途中に現れる情報 (キーワード) と，最終的に利用者が取得すべき情報 (文章) が乖離していた．これに対し，本論文で検討する上述の方法では，両者に境界が無いので，最終段階という概念が無く，要約文章を読み進めている最中に利用者の情報要求が満足されれば，そこで文章を読むことをやめるだけでよい．利用者は要約文章を読み進め，重要箇所に印をつける作業をするだけである．そのため利用者は文書を読むことに専念でき，効率良く情報検索を行なえると考えられる．

また，要約文書を読み進めるにつれて利用者の興味が次第に詳細化したり移り代わっていくことも有り得るので，これにも対応したい．そのために，適合性フィードバックを利用して利用者の興味に適合した関連文書の再検索を行ない，それらの要約を提示することを提案する．これを推薦情報要約と呼ぶことにする．ここで，主たる要約文章と上記の推薦情報要約文章は分離されて提示される点，ならびに，利用者はその両者に注目箇所の印を付与できる点に注意されたい．これにより，利用者が行っている情報の絞り込みを妨げず，かつ，利用者の意思によって必要に応じて推薦情報を取り込める．この過程も図 1に示した．

3.2 対話型要約システムの実装

前節で述べた基本的な考え方に基づき，次の手順に従う対話型要約システムを実装した．図 2にシステムの表示例を示す．

1. 文書検索のために利用者が入力した任意のキーワード列から，検索ベクトルを生成し，文書データベース中の関連文書を検索する．状況により，検索結果の中から要約したい文書群を利用者に選択してもらい，初期文書集合とする．

2. 与えられた文書集合の要約 (主要約文章) を生成し，表示する．要約処理過程では文書集合のクラスタリングが行なわれる． (Scatter 過程に相当)

3. 表示された要約文章 (主要約文章，ならびに，二巡目以降は推薦情報要約文章も) の中から，利用者は興味にしたがって自由に文字列を選択する．(Gather 過程の一部に相当)

4. 利用者の選んだそれぞれの文字列の属するクラスタをまとめて次の要約処理の対象とする．なお，状況に応じて利用者は，それぞれの文字列の由来する文書のみを次の要約対象とすることができる．(Gather 過程の一部に相当)

5. 2. で用いた文書集合を，4. で選択された関連文書集合と，それ以外の文書集合 (非関連文書集合に対応する) に分割する．これらを用いて，適合性フィードバック手法により検索ベクトルを更新し，関連推薦文書の検索を行なう．さらに，それらの要約文章 (推薦情報要約文章) を生成する．

6. 2. に戻り，主要約文章を生成し，上記の推薦情報要約文章とともに表示する．

手順 2. から 6. までを繰り返すことにより，利用者の興味に応じて文書集合が絞り込まれ，対応する要約文章が生成される．また，利用者が推薦情報要約から文字列を選択した場合はそれの由来する文書 (群) が要約対象に追加され，興味の移り変わりが主要約文章に反映される．

なお，本システムで利用している複数文書要約手法は，まず tf・idf 法による語の重要度に基づき文重要度を各文に付与し， MMR を拡張した MMI-MS[7] によって冗長性排除を行なう重要文抽出手法である．

まず，文 s の重要度 Imp_s は以下の式で計算している．

$$Imp_s = \frac{\sum_{h \in s} TF_s(h) \times TF_d(h) \times BIAS(h) \times IDF(h)}{N_w} \quad (1)$$

ここで h は文 s 中の語， $TF_s(h)$ は文 s 中に含まれる語 h の数， $TF_d(h)$ は要約対象文書集合中に含まれる語 h の数を文書集合中の語数で正規化したもの， $IDF(h)$ は文書データベースにおける h の IDF 値， N_w は s 中の語数である． $BIAS(h)$ は利用者が選択した文字列に応じて， h に対して計算されるバイアス値である．これは 3.4節で説明する．

さらに， MMI-MS を用いて冗長性の制御を行ない，要約文章として提示する文を選ぶ． MMI-MS は次式で定義される．

$$MMI\text{-}MS(SS, A) = Arg \max_{s_i \in SS \setminus A} [\lambda Imp_{s_i} - (1-\lambda) \max_{s_j \in A} Sim_s(s_i, s_j)] \quad (2)$$

ここで SS は要約対象文書集合中のすべての文集合， Sim_s は文間類似度であり， λ は重要度と類似度のどちらを重視するかを決めるパラメータである．また， A は要約文章として既に採用された文集合であり，初期は空集合である．この式を繰り返し適用することで，文の重要度と要約の冗長性を同時に考慮しながら要約文章として採用する文集合を選ぶ．なお，推薦情報要約生成の過程では， A は空集合ではなく，主要約文章の文集合になっている．これによって主要約との冗長性を制御しながら推薦情報要約を生成する．

選ばれた重要文は由来する文書の日付に基づき次のように整列される．まず，要約対象となった原文書集合を単一リンク法によりクラスタリングをし，類似文書のクラスタ群を得る．各クラスタ内で各文書を作成日時に従っ

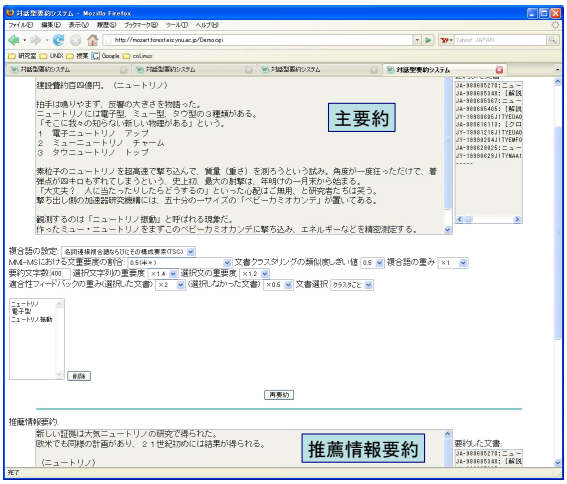

図 2: 対話型要約システムの動作例

て整列する．ついで，クラスタ内の一番古い文書の日付に従って，クラスタを整列する．こうして，原文書の順序を得る．この原文書の順序に従って，抽出された各重要文を整列する．

3.3 対話を行なうインターフェース

提案手法では，利用者が要約文章に印をつける作業をマウスで行なえるようになっている．表示された要約文章中の文字列をマウスでドラッグして選択すると，その文字列に印をつけることができる．また，要約の中から必要な原文書が見つかった場合は，その原文書を参照することもできる．

利用者は基本的に興味のある表現を選んで次の要約に進む操作をするのみであり，その選択した表現の含まれるクラスタなどは見えないことに注意されたい．本手法は，利用者に文章を読むこと以外の負担をなるべくかけずに，利用者の興味をフィードバックすることを目的とするものである．

3.4 利用者の興味を考慮した要約の生成

利用者は提示された要約文章から任意の文字列を選ぶ．これにより次のような情報が得られる．

- 選択された文字列
- 選択された文字列を含む文
- 選択された文字列を含む文書
- 選択された文字列を含むクラスタ

利用者が選択した文字列を含む文書およびクラスタは，Gather 過程において次の要約対象となる文書集合を構成する際に用いられるが，提案手法ではより粒度の細かい情報まで得ることができる．そこで，重要文抽出による要約生成の際に，利用者の選択した文字列に含まれる語の重みを高くすることで，その文字列が重要視されるようにした．また，選択された文字列を含む文中に現れる他の語についても，選択した文字列と関連する語であるので重みを高くした．これらは $BIAS(h)$ の値として実現される．しかし文や文書については，その重要度は単語の重要度に還元されると考え，文や文書単位で重みを高くすることは行なわなかった．

3.5 推薦情報要約

Scatter/Gather 法による文書集合の絞り込みでは，文書集合は少なくなる一方で増えることはない．しかし，要約文書を読み進めるにつれて利用者の興味が次第に詳細化したり移り代わっていくことも有り得るので，文書集合中にない文書の情報を提示することも利用者の情報要求に対して有効に働くと考えられる．そこで，適合性フィードバックを利用して検索ベクトルを更新し，文書の再検索を行ない，その要約を推薦情報要約として提示することを提案する．要約に用いた文書集合の中で，利用者が選択した文字列を含むクラスタを利用者の興味に適合したクラスタとし，それ以外のクラスタを適合しなかったクラスタとして，適合性フィードバックにより検索ベクトル q を更新する．実装したシステムでは以下の Rocchio の式を用いた．

$$q' = q + \frac{\alpha}{|D_R|}\sum_{d_i \in D_R} d_i - \frac{\beta}{|D_N|}\sum_{d_j \in D_N} d_j \qquad (3)$$

ここで D_R, D_N はそれぞれ，適合 / 不適合と判断されたクラスタ群を併合した文書の集合である．α, β はそれぞれの文書集合をどの程度重要視するかを決める定数で，0 以上の値をとる．

更新された検索ベクトル q' を用いた文書検索の結果を対象として，主要約と同様の要約処理を行ない，推薦情報要約を生成する．これによって主要約の文書集合とは違う観点から利用者の興味と関連のある要約を提示でき，利用者の興味の変化に対応できると考えられる．

4 評価実験

4.1 実験方法

提案手法における文書集合の絞り込みと，推薦情報要約の有効性を評価するために評価実験を行なった．文書データベースとして毎日，読売新聞の 98，99 年の新聞記事を用い，それぞれの手法を比較するために以下の 3 つのシステムを用意した．

- 任意のキーワードで文書検索を行ない，検索結果から記事の見出しを見て要約したい文書群を選び，要約を行なう．(baseline)
- baseline に加え，表示された主要約文章から任意の文字列を選択でき，絞り込み要約を行なえる．(提案手法 A)
- 提案手法 A に加え，絞り込み要約後は推薦文章要約も表示され，どちらからも文字列を選択し，再要約を行なえる．(提案手法 B)

なお，いずれのシステムも任意の時点で文書検索からやり直すことができる．実験は，あるトピックに関する複数の質問について，人間がその答を要約システムを用いて探すというタスクで行ない，その正答率と被験者の行なった操作の回数を評価の対象とした．被験者の答える質問には NTCIR4 TSC3 Formal Run[8] において要約生成の一助として与えられていた質問文のリストを用

いた．同 Formal Run は 30 トピックから構成されていたが，そのうち無作為に抽出した 3 つのトピックについて実験を行なった．

被験者は工学を専攻する 8 人の学生で，それぞれ同一のトピックを 2 回ずつ別々のシステムで解き，計 6 回作業を行なってもらった．また，「同じトピックの問題を解いた 2 つのシステムのうち，どちらが情報を得やすかったか」という問いにより優劣を判定してもらった．さらに，自由記述形式で，各トピックについてそれぞれのシステムのうまくいった点，いかなかった点および感想を事後に記述してもらった．なお 1 つのトピックを解く時間は 20 分に制限した．適合性フィードバックの定数は $\alpha = 2.0, \beta = 0.5$ とし，選択された文字列に含まれる語の重みは $BIAS(h) = 1.4$，その文字列を含む文中の単語は $BIAS(h) = 1.2$ として実験を行なった．

4.2 実験結果

被験者の正答率，および操作回数として記録したマウスクリック数，文書検索の回数，要約・再要約の回数結果を被験者の数で平均したものを表 1に示す．なお，主観評価の欄には，2 つのシステムを比較してこちらがより情報を得やすいと判断した被験者の人数を記入した．

表 1: 評価実験における正答率，主観評価と操作回数

	base vs 提案 A		提案 A vs 提案 B		base vs 提案 B	
システム	base	提案 A	提案 A	提案 B	base	提案 B
正答率	71.5%	65.1%	73.8%	63.7%	66.9%	68.6%
主観評価	3 人	5 人	4 人	4 人	1.5 人	6.5 人
クリック数	152	136	125	121	129	146
検索回数	12.3	9.1	7.6	8.9	10.1	9.1
要約回数	16.4	10.8	11.3	8.6	14.0	11.8
再要約回数	0	5.6	4.6	7.2	0	5.4
原文書表示回数	3.9	1.4	1.4	2.6	3.1	2.6

baseline と提案手法 A の比較では，正答率は baseline の方が高いものの，被験者の主観による評価では提案手法 A が上回り，操作回数も提案手法 A の方が少なく，より効率的に情報が得られていると言える．提案手法 A と B の比較では正答率は提案手法 A の方が高く，その他の要素は両者とも大きな差はない． baseline と提案手法 B の比較では，被験者の主観による評価は提案手法 B が上回るが，操作回数はむしろ提案手法 B の方が多いと言える．

正答率については各システム間でそれほど大きな差はなく，特定の情報を見つけるというタスクにおいてはどれも性能は変わらないと言える．これは， baseline の「検索結果から記事の見出しを見て要約したい文書群を選び，要約を行なう」という操作が，必要な文書を人手で絞り込む作業に相当し，これだけで実験のタスクを行なうのに十分な性能を持っていたことが原因と考えられる．

提案手法の目的は，利用者が自分の興味のある情報をどんどん読み進めていけることであるので，その観点から考察を行なう．まず baseline と提案手法 B で被験者の主観評価に差があるが，その理由として被験者は，提案手法 B で提示される推薦情報要約で関連する答が得られたことを多く挙げていた．これはすなわち，推薦情報要約が利用者の興味に関連する情報をうまく提示できていたことを示す． baseline ではすべて手動で答を探さなければいけないが，提案手法 B では自動で答が見つかる場合もあったことが評価されたとも考えられる．

また，原文書の表示回数を見ると， baseline と提案手法 A， baseline と提案手法 B ではどちらも baseline の方が表示回数が多い．表示回数の少ない提案手法 A， B はそれだけ提示した要約文章に必要な情報が含まれていたと考えられる．よって文書集合を絞り込んだ再要約および推薦情報要約は，利用者の興味に応じた情報提示に効果があったと考えられる．しかし一方で，文書集合を絞り込んでも要約文章が変わらなかったという報告が多くあったので，クラスタ数が多くなるクラスタリング手法を採用するとともに，再要約の際にはすでに表示した要約文章との重なりを考慮する必要があると考えられる．

5 まとめ

本論文では Scatter/Gather 法を要約提示の観点から捉え直す事により，対話的なシステムで利用者の興味を反映した要約を生成する手法を提案した．また，同手法について人手による評価を行なった．

今後は評価実験の結果をもとに，インターフェースの改良を行なう予定である．また，提案手法における Scatter 過程は Stein ら [9] や Radev ら [10] の提唱する，文書クラスタリングに基づく複数文書要約手法と直接対応づけられる．そのため，これらは提案手法に適した要約手法と考えられるので，今後その実装について検討したい．

参考文献

[1] Anastasios Tombros and Mark Sanderson. Advantages of query biased summaries in information retrieval. In *SIGIR*, pp. 2–10. ACM, 1998.

[2] 平尾努, 佐々木裕, 磯崎秀樹. 質問に適応した文書要約手法とその評価. 情報処理学会論文誌, Vol. 42, No. 9, pp. 2259–2269, 2001.

[3] Tatsunori Mori, Masanori Nozawa, and Yoshiaki Asada. Multi-answer-focused multi-document summarization using a question-answering engine. *ACM Trans. Asian Lang. Inf. Process.*, Vol. 4, No. 3, pp. 305–320, 2005.

[4] Katashi Nagao and Kôiti Hasida. Automatic text summarization based on the global document annotation. In *COLING-ACL*, pp. 917–921, 1998.

[5] Anton Leuski, Chin-Yew Lin, and Eduard Hovy. iNeATS:interactive multi-document summarization. In *Proceedings of 41st Annual Meeting of the Association for Computational Linguistics(ACL-03)*, pp. 125–128, 2003.

[6] Douglas R. Cutting, Jan O. Pedersen, David R. Karger, and John W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Nicholas J. Belkin, Peter Ingwersen, and Annelise Mark Pejtersen, editors, *SIGIR*, pp. 318–329. ACM, 1992.

[7] 佐々木拓郎. 情報利得比に基づく語の重要度と mmr の統合による複数文書要約. Master's thesis, 横浜国立大学大学院 環境情報学府情報メディア環境学専攻 情報メディア学コース, 2003.

[8] TSC 実行委員会. NTCIR-4 テキスト自動要約タスク (automatic text summarization task)/TSC-3(text summarization challenge - 3). http://lr-www.pi.titech.ac.jp/tsc/tsc3.html, 2003.

[9] Gees C. Stein, Tomek Strzalkowski, and G. Bowden Wise. Interactive, text-based summarization of multiple documents. *Computational Intelligence*, Vol. 16, No. 4, pp. 606–613, 2000.

[10] Dragomir R. Radev, Hongyan Jing, Magorzata Sty, and Daniel Tam. Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, Vol. 40, No. 6, pp. 919–938, 2004.