クエリに基づくインタラクティブな文書要約システムの検討

A Study on Interactive Summarization System of Documents Based on Queries

田村一樹[1*] 吉川大弘[2] 古橋武[2]
Kazuki Tamura[1], Tomohiro Yoshikawa[2], and Takeshi Furuhashi[2]

[1] 名古屋大学工学部
[1] School of Engineering Nagoya University
[2] 名古屋大学工学研究科
[2] Graduate School of Engineering Nagoya University

Abstract: Recently, digitization of paper-based documents is rapidly advanced. In near future, it is expected that we can easily read various kinds of electronic books on the Internet. However, it becomes difficult for us to choose suitable ones from the huge amount of documents in a short time. Then it will need the support system to select the demanded books. The summarization of a book will be one of the effective ways to understand their contents. Though there are a lot of researches for the summarization, most of them focus on the quantification of the importance of each sentence and the automatical extraction of the important sentences. This paper aims to construct the interactive summarization system using queries. This system repeats the input of query by a user and the extraction and presentation of important sentences with the query. It is expected that a user can understand the contents of a book through this interaction. This paper studies the effectiveness of the interactive summarization through an examination.

1 はじめに

近年，紙媒体で出版されてきた書籍の電子化が急速に進められており，近い将来，我々が気軽に読むことのできる電子書籍の数は膨大になると予想される．この電子書籍はこれまで，主に PC 向けと携帯電話向けの市場を中心に成長してきたが，今後は，昨今のタブレット端末・スマートフォンの普及により，これらの市場を加えることで電子書籍の利用者はさらに増大すると思われる．インプレス R&D の「電子書籍ビジネス調査報告書 2011」によると，2010 年度の電子書籍市場規模は 650 億円と推計され，さらに 2015 年には約 3.1 倍の 2000 億円程度に成長すると予測されている [1].

電子化により，これまでの紙媒体ではできなかった書籍の利用方法が考えられる．その一つが，自動要約された要約文によって内容を把握するという方法である．従来は，内容を把握するには実際に書籍を通読するしかなく，時間と労力の負担が読者にかかっていた．特に自己啓発書のような書籍では，読者の興味は文脈などの文学的な側面ではなく，実用的な記述がされている箇所にあると考えられる．それらを要約としてまとめて提示することで，書籍を通読することなく内容の把握をすることができ，読者の負担の軽減につながると考えられる．

本稿では，自動的な処理によって一意に要約を生成・提示するのではなく，ユーザが積極的に介入し，インタラクティブな操作により文書の要約を行うシステムの構築を目指す．本システムでは，ユーザは興味・関心を持った語句をクエリとして設定する．語句可視化ツール HK Graph (Hierarchical Keyword Graph) を用いて文書中のキーワードをユーザに提示し，クエリの選択支援を行う．システムは選択されたクエリに基づいて各文の重要度計算を行い，重要度の高い文を要約としてユーザに提示する．ユーザは提示された要約を基に，さらに新たなクエリを設定し，これによりシステムは新たな要約を再提示する．このインタラクションによって，ユーザの求める様々な視点からの要約が逐次生成され，文書の内容が把握可能となることが期待される．

*連絡先： 名古屋大学工学部電気電子・情報工学科
名古屋市千種区不老町
E-mail: tamura@cmplx.cse.nagoya-u.ac.jp

2　従来研究

要約には大きく“ 抜粋 (extract) ”と“ アブストラクト (abstract) ”の 2 つのアプローチが存在する [2]．抜粋とは，文書中の単語や文を抽出し，それらを並べることで要約とする手法で，アブストラクトは言い換えや結合などの自由作文的な操作を加える手法である．従来の要約手法では，自然言語処理によって得られた情報から，統計的手法などによって各文の重要度を計算し，重要度の高い文章を抜粋して要約とする方法が多く取られてきた [3]．他に，テキスト構造を用いるもの [4]，機械学習を用いるもの [5] などが研究されているが，いずれも人が作成する要約には及んでいない．

また，ユーザの興味・関心を反映する要約手法として，query-biased な要約手法がある [6]．query-biased な手法はこれまで，複数文書から目的の文書を探すための情報検索の手法として用いられてきた [7] が，近年，要約文を提示することで文書の内容把握支援を行う研究も報告され始めている [8]．これはユーザがクエリを設定することで，文書全体を抽象的に説明する要約ではなく，設定されたクエリに特化した要約を生成する手法である．この方法では，クエリに関係する単語や文の重要度を高くし，その重要度が上位となる文を抜粋することで要約とするアプローチが多い．しかし，文書が複数の内容によって構成されている場合など，一度クエリを設定し，要約文を提示するだけでは，内容の理解が十分にできない場合がある．そこで本研究では，一意に要約を生成・提示するのではなく，ユーザとのインタラクションによって何回か要約を生成するシステムについて検討する．

3　提案システム

提案システムでは，初めに CaboCha[9] を用いて形態素解析を行う．形態素解析の結果に基づき，クエリ候補となるキーワードを HK Graph[10] を用いてユーザに提示する．ユーザは自らの興味・関心と，HK Graph からの情報を基にクエリを設定する．システムはクエリに応じて，文書中の全単語に対して重要度計算を行う．その後，各単語の重要度から各文の重要度を計算する．ユーザが指定した要約率に達するまで，重要度が上位の文から順に抜粋し，抜粋された文を本文での出現順に並べ，要約としてユーザに提示する．その要約文では満足に理解できなかった場合，ユーザは提示された要約，もしくは HK Graph の情報を基に新たなクエリを設定し，システムを再実行する．システムはこれまでに設定されたクエリの情報から，新たに単語・文の重要度計算を行い，要約を再提示する．ユーザは文書に対して満足な理解が得られるまで，システムとのインタラクションを続ける．提案システムのイメージを図 1 に示す．

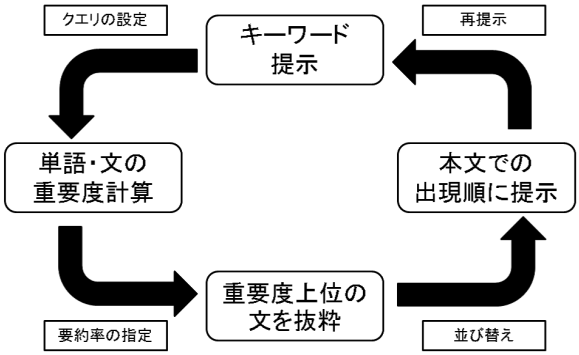


図 1: 提案システム

3.1　キーワード提示

キーワード提示には，文書中の語句可視化ツールである HK Graph を用いる．HK Graph とは，テキストマイニングを目的に，筆者らにより開発された解析ツールである．HK Graph では，入力された文書に対し，CaboCha を用いて形態素解析・係り受け解析を行う．解析結果から，各単語について頻度などの重要度を計算し，グラフ構造でキーワードを提示する．また HK Graph では，共起度や係り受け情報などを基に，階層的に単語をつないでいくことも可能である．HK Graph そのものを用いた文書の内容把握も可能ではあるが，本稿では，HK Graph によってキーワードを提示し，文書の特徴を示すことによる，ユーザのクエリ選択支援方法として用いる．HK Graph でのキーワード提示の様子を図 2 に示す．

図 2: HK Graph によるキーワードの提示

3.2 重要度計算

ユーザの選択したクエリを基に，各単語 (本稿では名詞のみとした) について重要度計算を行う．単語の重要度には，文書中での出現頻度と，各文中でのクエリとの共起度を用いる．

文書中に出現する全単語 (名詞) 数を N_w とし，ある単語 w_i について本文中での出現回数を $N_a(w_i)$ とすると，w_i の頻度スコア $tfScore(w_i)$ は式 (1) で計算される．

$$tfScore(w_i) = \frac{N_a(w_i)}{\sum_{j=1}^{N_w} N_a(w_j)} \tag{1}$$

また，m 回目の試行において設定されたクエリを q_m とし，w_i の q_m との共起回数を $N_c(w_i)$ とすると，w_i の共起スコア $coScore(w_i)$ は式 (2) で計算される．

$$coScore(w_i) = \frac{N_c(w_i)}{\sum_{j=1}^{N_w} N_c(w_j)} \tag{2}$$

これらから単語 w_i の重要度 $I_w(w_i)$ は式 (3) で計算される．

$$I_w(w_i) = tfScore(w_i) + coScore(w_i) \tag{3}$$

これら各単語の重要度に基づき，各文に対する重要度を求める．文 S_i 中に出現する単語の集合を $W = \{w_1, w_2, \ldots, w_{N_{S_i}}\}$ とすると，S_i の重要度 $I_s(S_i)$ は式 (4) で求められる．ただし $N_{\bar{Q}}(S_i)$ は，W のうちクエリ集合 $Q = \{q_1, q_2, \ldots, q_m\}$ に含まれない単語の数である．式 (4) において Q を除いたのは，あるクエリで提示された要約文を見て，ユーザが次のクエリを設定することを想定した場合，共起スコアの高い単語があることで重要文が決まるため，クエリ同士が共起し合うことで，過去に提示された文と同じ文が重要文として抜粋されてしまうことを避けるためである．例えば，ある文書で「生活のリズム」という表現が頻出している場合，クエリ「生活」で重要となる文は，それら「生活のリズム」の表現が使われている文であり，その要約を見てクエリ「リズム」を設定しても，重要度が上位にくるのは同じ文となる．同じ文を提示することを避ける処理を行ったとしても，それは「生活」のクエリで提示する文の数を増やした場合と等しくなってしまう．

$$I_s(S_i) = \frac{\sum_{j=1}^{N_{S_i}} I_w(w_j \notin Q)}{N_{\bar{Q}}(S_i)} \tag{4}$$

3.3 抜粋・提示

3.2 で行った各文に対する重要度に基づき，重要度上位の文から，ユーザが指定した要約率に達するまで文の抜粋を行う．このとき，それまでの試行で既に提示された文は提示しない．文書の全文章数を N_s，要約率を r[%] とすると，提示する要約文の数は $N_s * r/100$ である．最後に，抜粋された文を本文での出現順に並べ，要約文としてユーザに提示する．

4 実験

4.1 比較実験

4.1.1 実験方法

インタラクティブな要約システムの有効性を検証するため，重要度の計算方法は同じとし，インタラクティブではない自動抜粋手法 2 つと比較実験を行った．自動的に抜粋を行う手法として，クエリに依存せず，一意に全体要約を提示する手法（手法 A）と，過去のクエリに依存せず，自動抽出された複数のクエリに対して抜粋を行う手法（手法 B）を用いた．手法 A では共起スコアは存在しないが，どちらも 3.2 で示した重要度を用いた．それぞれ抜粋・提示される文の数を等しくして比較を行った．

手法 B のクエリは，文書中の出現回数が多い名詞から順に設定した．一方提案手法のクエリは，実際にユーザが使用する場合を想定し，以下のルールで設定する．

1. 最初に設定するクエリは文書中の出現回数が最多の語とする．
2. 要約が提示されたら，その要約の中で出現回数が最多の語を次のクエリに設定する．
3. 2. で最多の語が 2 語以上ある場合，それまでの出現回数の累計が最多の語をクエリに設定する．

このルールは，ユーザは提示された要約を読み，そこで多く使われている語にさらに興味を持つという仮定に基づいている．

実験には，「脳が冴える 15 の習慣―記憶・集中・思考力を高める」[11] の第一章のみを用いた．文の数は 147 文，全 5700 字からなる文書であり，名詞は全 314 種類，延べ出現回数は 669 回であった．HK Graph では，文書での出現回数が多い順に名詞 30 語句を提示した．手法 B および提案手法では，一試行で 3 文 (要約率 2.0%) を提示し，順にクエリを設定し，手法 A は要約率を 2.0% ,4.0% ,... と変化させた．提示された文の数に応じて，4.1.2 で示す F 値で比較・評価を行った．それぞれの手法について，合計 21 文の提示となるまで実験を行った．

4.1.2 評価指標

評価を行うにあたり，用いた文書の末尾にある筆者のまとめ文を取り除き，そのまとめ文の情報を基に，6 項目の正解情報を定義した．文書中において，この 6 項目の正解情報のいずれかが読み取れる文を事前に正解文 (17 文) として定義した．

生成された要約文に対して，適合率 (*precision*)，再現率 (*recall*) を求め，情報検索の性能を表す指標として多く用いられている F 値 (*F-measure*) を用いて比較・評価を行った．図 3 に，適合率・再現率についてのイメージを示し，F 値を式 (5) に示す．

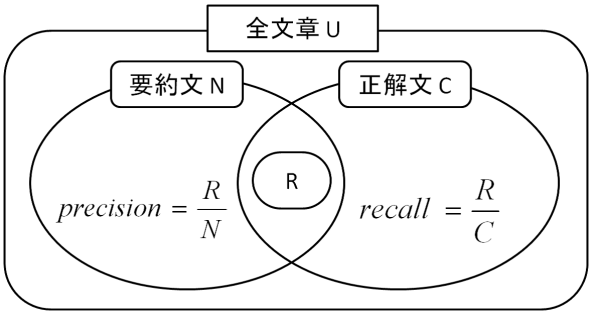


図 3: 適合率・再現率

$$F\text{-}measure = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{5}$$

4.1.3 結果および考察

得られた結果を表 1 に，表 1(a)(b)(c) それぞれをグラフで表したものを図 4 に示す．

結果から，提案手法の F 値が，総じて比較手法より高くなっていることがわかる．また内訳を見ると，特に適合率が高く，これが F 値を高くしている要因であることがわかる．実際に要約文を確認すると，手法 A ではクエリの情報を含んでいないため，視点の定まらない要約が生成されていた．一方で手法 B は，クエリの情報を含んでいるため，クエリに関連した視点での要約は生成されていた．しかし，過去のクエリの情報を用いていないことで，冗長な要約が生成される場合がみられた．提案手法では，過去に設定したクエリをすでに理解した情報として用いているため，幅広く要約が生成されていた．

4.2 ユーザ実験

4.2.1 実験方法

提案システムのクエリの設定順について，4.1.1 で示した方法が，どの程度実際にユーザが使用する場合と合致しているかを検証するため，ユーザ実験を行った．実験データや条件，および評価指標は 4.1 と同じとし，クエリのみ，ユーザに自由に設定してもらった．被験者は，事前にこの文書を読んだことがない大学生 3 名で，提案システムを用いて内容の把握に努めてもらった．クエリの設定は 7 回まで，つまり合計 21 文までの提示とした．

4.2.2 結果および考察

被験者をそれぞれユーザ a，ユーザ b，ユーザ c とし，仮想ユーザは 4.1.1 で定めた手順でクエリを設定するとした．得られた結果を表 2 に，表 2 の F 値をグラフで表したものを図 5 にそれぞれ示す．

ユーザ a およびユーザ b は，概ね仮想ユーザと近い F 値を持つ結果となった．一方，ユーザ c は仮想ユーザと比較して低い F 値となっている．ユーザ c のクエリの設定順を確認すると，提示された要約には含まれない語を次のクエリに設定することが複数回あった．つまり，ユーザ c は生成された要約に従って内容を読み進めるのではなく，全体を広く把握できるよう，様々な視点から読み進めようとしたと考えられる．また，ユーザ c は結果的に，正解情報には含まれない内容に関係したクエリを多く設定しており，正解文があまり抜き出せなかったと考えられる．一方でユーザ a とユーザ b は，提示された要約に従って正解情報に含まれるような内容を読み進めており，4.1.1 で仮定したクエリの設定順は実際のユーザの使い方と大きくはずれていないことが確認された．

本ユーザ実験により，実際に使用する際のクエリ設定の仕方次第で大きく要約結果が変わることがわかった．提案システムでは，ユーザの興味・関心に沿うような読み進め方は可能だが，文書の主題となるような内容を網羅的に把握するためには，クエリの設定の仕方に大きく依存しており，クエリや要約と関連の深い語を強調・推薦するなど，次のクエリの選択に何らかの補助が必要であると考えられる．

5 まとめ

本稿では，ユーザが積極的に介入し，インタラクティブな操作により文書の要約を行うことで，文書の内容の把握を支援するシステムについて検討を行った．実際の文書要約に対する比較実験を行い，自動的に抜粋

を行うシステムと比較して，提案システムの適合率・再現率が全体的に高く，情報の質が高い要約が生成されることを示した．また，ユーザ実験により，ユーザによってクエリの設定の仕方が異なり，それが要約の精度に少なからず影響を与えることを示した．今後の課題として，頻度や共起などの表層的な情報だけではなく，意味や文間の関係などを考慮した単語・文の重要度計算の方法を導入することや，ユーザへの効果的なクエリ候補の提示方法に対する検討などが挙げられる．

表 1: 各手法の評価指標

(a) 適合率

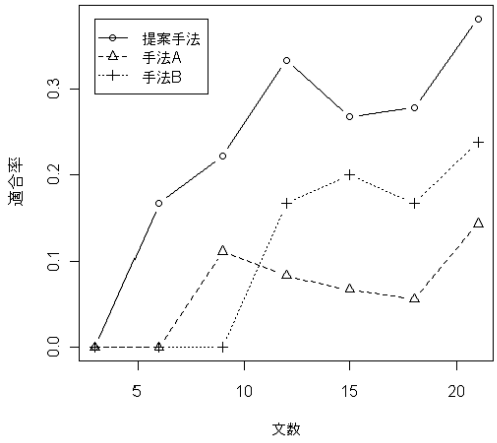
抜粋文数	提案手法	手法 A	手法 B
3	0	0	0
6	0.167	0	0
9	0.222	0.111	0
12	0.333	0.083	0.167
15	0.267	0.067	0.200
18	0.278	0.056	0.167
21	0.381	0.143	0.238

(b) 再現率

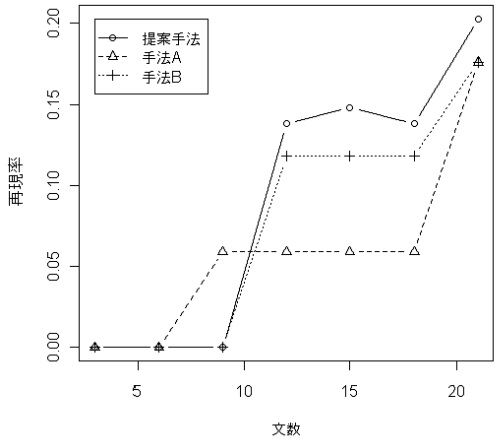
抜粋文数	提案手法	手法 A	手法 B
3	0	0	0
6	0	0	0
9	0	0.059	0
12	0.138	0.059	0.118
15	0.148	0.059	0.118
18	0.138	0.059	0.118
21	0.203	0.176	0.176

(c) F 値

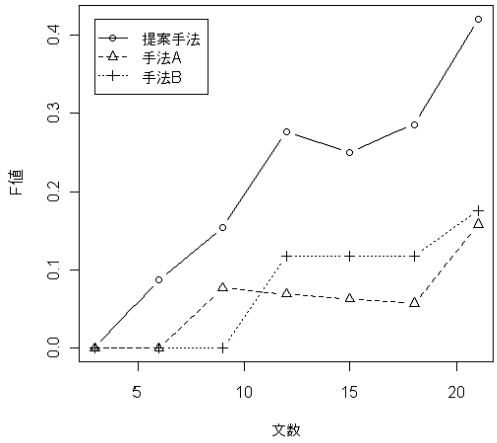
抜粋文数	提案手法	手法 A	手法 B
3	0	0	0
6	0.087	0	0
9	0.154	0.077	0
12	0.276	0.069	0.118
15	0.250	0.063	0.118
18	0.286	0.057	0.118
21	0.421	0.158	0.176



(a) 適合率



(b) 再現率



(c) F 値

図 4: 評価指標による手法の比較

謝辞

本研究は，文部科学省科学研究費（基盤研究 (C), No.22500088）の補助を得て遂行された．

表 2: 各ユーザにおける評価指標

(a) 適合率

文数	ユーザ a	ユーザ b	ユーザ c	仮想ユーザ
3	0.333	0	0	0
6	0.333	0.167	0	0.167
9	0.444	0.111	0.111	0.222
12	0.333	0.167	0.167	0.333
15	0.267	0.200	0.133	0.267
18	0.278	0.222	0.167	0.278
21	0.333	0.333	0.143	0.381

(b) 再現率

文数	ユーザ a	ユーザ b	ユーザ c	仮想ユーザ
3	0.059	0	0	0
6	0.118	0.059	0	0
9	0.235	0.059	0.059	0
12	0.235	0.118	0.118	0.138
15	0.235	0.176	0.118	0.148
18	0.294	0.235	0.176	0.138
21	0.412	0.412	0.176	0.203

(c) F 値

文数	ユーザ a	ユーザ b	ユーザ c	仮想ユーザ
3	0.100	0	0	0
6	0.174	0.087	0	0.087
9	0.308	0.077	0.077	0.154
12	0.276	0.138	0.138	0.276
15	0.250	0.188	0.125	0.250
18	0.286	0.229	0.171	0.286
21	0.368	0.368	0.158	0.421

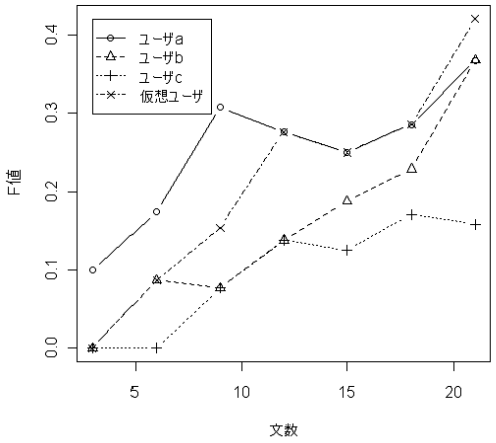


図 5: 評価指標のグラフ (F 値)

参考文献

[1] 電子書籍ビジネス調査報告書 2011, 株式会社 インプレス R&D (2011)

[2] Mani, I.: Automatic Summarization, John Benjamin s Publishing Company (2001)

[3] Edmundson, H.P.: New methods in automatic abstracting, *Journal of ACM*, Vol. 16, No. 2, pp. 264–285 (1969)

[4] 綾 聡平, 松尾 豊, 岡崎 直観, 橋田 浩一, 石塚 満: 修辞構造のアノテーションに基づく要約生成, 人工知能学会論文誌, Vol. 20, No. 3, pp. 149–158 (2005)

[5] 鈴木 大介, 内海 彰: Support Vector Machine を用いた文書の重要文節抽出：要約文生成に向けて, 人工知能学会論文誌, Vol. 21, No. 4, pp. 330–339 (2006)

[6] Tombros, A., Sanderson, M.: Advantages of query biased summaries in information retrieval, *Proc. of the 21st Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, pp. 2–10 (1998)

[7] 桜井 俊彦, 内海 彰: 情報検索のためのクエリに基づく文書自動要約, 言語処理学会年次大会発表論文集, Vol. 10, pp. A10C2–03 (2004)

[8] 大谷 力, 織田 泰司, 内田 佳孝, 文 景厚, 古江 敏彦, 吉江 修: パラグラフ間の差異を考慮した query-biased な要約手法, 電気学会論文誌 *C*, Vol. 130, No. 12, pp. 2256–2265 (2010)

[9] 工藤 拓, 松本 祐治: チャンキングの段階適用による係り受け解析, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842 (2002)

[10] 岡部 貴博, 吉川 大弘, 古橋 武: メタデータと語句の共起情報を利用したインシデントレポート解析支援システムの提案, 日本知能ファジィ学会誌, Vol. 18, No. 5, pp. 689–700 (2006)

[11] 築山 節: 脳が冴える 15 の習慣—記憶・集中・思考力を高める, 生活人新書, 日本放送出版協会 (2006)