統計情報を用いた話題特定と文脈の再構築による複数テキスト要約

Multiple-text Summarization Method based on Topic Identification and Context Restructuring.

福原　知宏 †・武田　英明 †・西田　豊明 ‡†
Tomohiro FUKUHARA†, Hideaki TAKEDA† and Toyoaki NISHIDA‡†
† 奈良先端科学技術大学院大学 情報科学研究科*
Graduate School of Information Science, Nara Institute of Science and Technology
‡ 東京大学大学院工学系研究科
School of Engineering, The University of Tokyo.

Abstract

We propose a multiple-text summarization method for finding an overview of a large body of texts using a topic identification method and context generation method. Topic identification method is based on the skewness and the kurtosis of a word frequency in a text. The skewness and the kurtosis represent a topical generality of a word. We identify a general or partial topic from the classified texts using the skewness and the kurtosis. Context generation method is based on the linkage of relevant sentences. We identify relevant sentences using theme and focus information of a sentence. We implemented proposed methods into Topic Showcase: a topic browsing system. Experiments show a capability of our method for finding overview of a large body of texts.

1　緒言

複数テキストを対象とした自動要約が求められている．今日，WWW ページに見られる大量のテキストを対象とした検索システムが利用されているが，検索結果自体が大量である点や，検索結果が分類されていないなどの問題から，効率的な情報収集は困難である．大量テキストからの効率的な情報収集には，テキスト集合に含まれる話題の把握による概要把握が重要である．

本稿では，概要把握を目的とした複数テキストの自動要約手法を提案する．概要とはテキスト集合を代表する文の集合である．本稿では，テキスト集合の特徴的な内容を示す話題の特定手法と，話題の内容を説明する要約の生成手法を提案する．

2　概要把握のための複数テキスト要約

本稿における**概要**とは，テキスト集合の内容を代表する文の集合である．内容の類似するテキスト集合では，テキスト間に共通する記述が存在する．この記述は，テキスト集合の特徴を知る上で有効である．特に，WWW ページのような大量のテキストを対象とした情報収集では，個々のテキストについての詳細な理解より，テキスト集合の概要をまとめて理解することが重要である [1]．本研究では，テキスト集合の全体的な理解を**概要把握**と呼ぶ．

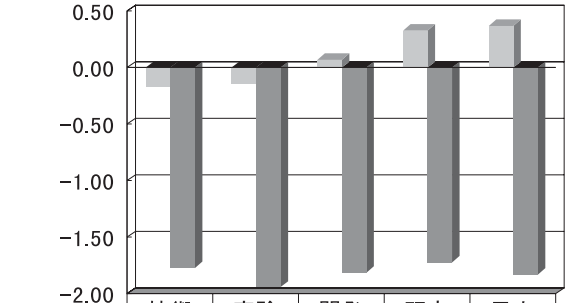
テキスト集合の概要把握に関する先行研究には，佐藤らの Netnews のダイジェスト自動生成 [2] や Hearst らのテキスト動的分類 [3] がある．

佐藤らの手法は，情報抽出に基づく概要把握支援である．この手法は，テキストの表層的特徴を用いて目的の情報を抽出し，ダイジェストとして自動編集する手法である．提案手法の利点は，一度，抽出規則を記述すれば，追加されるテキストからも同じように情報を抽出できる点である．一方，抽出規則の設定が困難な点や対象テキストが限定されるなどの問題がある．

Hearst らの手法は，テキスト自動分類に基づく概要把握支援である．Hearst らは，ユーザの選択したテキスト集合を動的に再分類するための手法を示した．提案手法を実装したシステム Scatter/Gather は，次のように動作する．(1) システムはテキスト集合を各クラスタに分類，各クラスタの内容を複数のキーワードで表す，(2) ユーザはキーワードを参照し，関心のあるクラスタを選択する．(3) システムはユーザの選択したクラスタに含まれるテキストを分類する．この手法の利点は，テキスト集合を動的に分類出来る点である．一方，クラスタの内容はキーワードで表されるため，クラスタの内容を十分に表現できないといった問題がある．

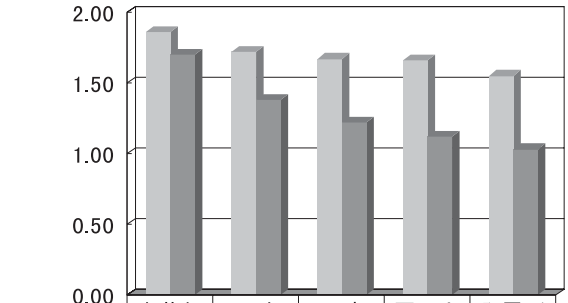
WWW のように雑多で大量のテキスト集合を対象と

*連絡先: 630-0101 生駒市高山町 8916-5 Tel (0743)79-5265 Fax (0743)79-5269 E-Mail: tomohi-f@is.aist-nara.ac.jp



	技術	実験	開発	研究	風力
skewness	−0.17	−0.14	0.07	0.33	0.37
kurtosis	−1.77	−1.94	−1.81	−1.72	−1.83

(a) Values of the skewness and the kurtosis of words in general topic



	水蒸気	石油	風速	原子力	発電所
skewness	1.86	1.72	1.67	1.66	1.55
kurtosis	1.70	1.38	1.22	1.12	1.03

(b) Values of the skewness and the kurtosis of words in partial topic

Fig. 1: Comparing the skewness and the kurtosis of words between partial topic and general topic

した概要把握では，対象テキストを限定しない話題特定手法が必要である．また，ユーザがテキスト集合について理解するため，特定した話題についての説明が必要である．

本研究では，(1) テキスト集合の統計的特徴を用いた話題特定手法と，(2) 話題に関連する文を組み合せ，話題を説明する要約を生成する手法を提案する．**話題**とは，相互に類似するテキスト集合の内容を代表する文である．本研究では，クラスタリングと単語の歪度・尖度を用い，対象テキストに依存しない話題特定手法を提案する．一方，**要約**とは話題に関連する文の集合である．本研究では，焦点-主題連鎖を用い，話題に関連する文を組み合わせる要約生成手法を提案する．

3　統計情報を用いた話題特定

クラスタリングと歪度・尖度を用いて話題を特定する．歪度・尖度は分布に関する統計値で，歪度は分布の歪みを，尖度は分布の尖りを表す指標である．歪度 α_3 と尖度 α_4 の定義を各々(1),(2) 式に示す [4]．

$$\alpha_3 = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^3}{s^3} \quad (1)$$

$$\alpha_4 = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^4}{s^4} - 3 \quad (2)$$

なお，n はデータ総数，$x_1, x_2, \cdots, x_n$ は各データ，$\bar{x}$ は平均，s は標準偏差である．

歪度と尖度がテキスト集合中の話題を示す例を Fig. 1 に示す．Fig. 1は，“風力発電” を含むテキスト集合中，全体的な話題を示す単語 (全体語)(Fig. 1(a)) と，部分的な話題を示す単語 (部分語)(Fig. 1(b)) の歪度 (skewness) と尖度 (kurtosis) を比較したグラフである．Fig. 1(b) の部分語は，Fig. 1(a) の全体語に比べ歪度・尖度ともに高い値をとることから，本研究では歪度と尖度を用いて全体的な話題・部分的な話題を特定する．

話題特定手順を次に示す．

1. テキスト集合をクラスタに分類．
2. 歪度・尖度から単語の評価値を計算．
3. 文に含まれる単語の評価値から，文の評価値を計算．
4. 文の評価値から部分話題・全体話題を特定．

クラスタリングには，最遠距離に基づく階層的クラスタリングを用いる．また，単語の評価値には歪度と尖度の線形和による次式を用いる．

$$\mathcal{P}(term_i) = w_1\alpha_3 + w_2\alpha_4 \quad (3)$$

なお，$term_i$ はテキスト集合中の単語 i，w_1, w_2 は各々，歪度 α_3 と尖度 α_4 に対する重みである．単語の全体語と部分語への分類は，(3) 式の評価値に閾値を設け，分類する．次に，テキスト中の各文について評価値を計算する．文の評価値は，文中に含まれる全体語もしくは部分語の評価値の総和である．全体話題は文に含まれる全体語の評価値の総和から，また，部分話題は部分語の総和から特定する．参考文献 [5] に詳細な手順を示す．

4　文脈の再構築による要約生成

文の主題と焦点情報を用いて要約の文脈を生成する．**主題**とは，ある一つの文中で中心的に記述される情報であり，**焦点**とは，後続の文の主題になり得る情報である．

本研究では，焦点-主題連鎖を用いて文脈を生成する．**焦点-主題連鎖**とは，先行文で取り上げた内容を後続の文で記述する談話構造である [6]．

焦点-主題連鎖のアルゴリズムを Table 1に示す．アルゴリズムの入力は要約の元となる話題 c_1，出力は要約

Table 1: Algorithm of summary generation procedure

inputs: $c_i \quad (i = 1)$
returns: $C = \{c_1, c_2, \cdots, c_N\}$

step 1 `if` $(i == N)$ 終了.

`while` $(i < N)$ 以下の step を繰り返す.

step 2 c_i の焦点集合 $F(c_i) = \{f_1^i, f_2^i, \cdots, f_m^i\}$ を求める．任意の焦点を選択し f_p とする.

step 3 f_p を主題に取る文集合 $S(f_p) = \{s_1^i, s_2^i, \cdots, s_n^i\}$ を求める．$S(f_p)$ から $s_q^i(\exists q\ max(\sum_{t \in s_q^i} \mathcal{P}(t)))$ を取り出す.

`if` $(s_q^i \in C)$ step 3 に戻る.

`else` $i = i + 1, c_i = s_q^i$. step 1 に戻る.

Table 2: An example of summary context.

1. 最近の**ソフトウエア開発**では エージェント利用 が新たな流れとなってきている。
2. ＮＴＴの開発した「追跡問題」モデルの**エージェント利用**は、それぞれの エージェント の働きを効率化し、全体の処理時間を短縮した点で前進といえる。
3. **エージェント**はそれぞれが独立し、自律判断機能を持つソフト。

の文脈 $C = \{c_1, c_2, \cdots, c_N\}$ である．アルゴリズム中で用いる関数は，文の焦点を求める関数 $F()$，焦点を主題に取る文の集合を求める関数 $S()$，要素の最大値を求める関数 $max()$ である.

アルゴリズムの手順は次の通り．まず，先行文 c_i に含まれる焦点集合 $F(c_i)$ から任意の焦点 f_p を選択する．次に，焦点 f_p を主題に取る文集合 $S(f_p)$ の内，最も多く全体語もしくは部分語を含む文 s_q を後続文 c_{i+1} として出力する．アルゴリズムは，要約の文数が N に達するまで繰り返す.

アルゴリズムを適用して生成した文脈の例を Table 2 に示す．表中の**太字**は文の主題を，下線 は焦点を表す.

5 概要把握支援システム: Topic Showcase

提案手法の実装として Topic Showcase を試作した．Topic Showcase は，テキスト集合の概要把握を支援するシステムである．ユーザはキーワード検索を行ない，クラスタリングされた検索結果から，各クラスタ及びテキスト集合全体の話題と要約を閲覧できる．システムの画面を Fig.2に示す．Fig.2中の文は，各クラスタの部分話題を表す．ユーザは，関心のある部分話題を選択することで，話題に応じた要約を閲覧できる.

システムの主な特徴は次の通り.

- キーワード検索を行い，検索結果をクラスタリングする.
- 各クラスタの部分話題，テキスト全体の全体話題を特定する.
- 特定した話題について要約を生成する.
- ユーザの選択したクラスタ数に応じ，動的に話題を更新する.
- ユーザの選択した話題に応じ，動的に要約を生成する.

Fig. 2: Screen image of Topic Showcase

Table 3: Ratio of an affirmative to negatives

Topic	Ratio $(Negatives/Affirmatives)$
Partial Topic (1)	0.75
Partial Topic (2)	0.64
Partial Topic (3)	0.36
General Topic	1.33

Topic Showcase では，ユーザがクラスタ数を設定し，話題の粒度を変更できる．また，ユーザはテキスト集合中の話題を選択し，話題に応じた要約を閲覧できる．このように，Topic Showcase ではユーザの観点に応じた概要把握が可能である.

システムは WWW ブラウザをインタフェースとし，Perl と CGI を用いて構築した．テキストデータには新聞記事[1]を用いた.

6 評価

提案手法の評価として，システムの特定した話題と要約についてアンケート調査を行なった．評価者は情報科

[1] 日本経済新聞 95 年度版 (本稿に掲載される新聞記事の著作権は日本経済新聞社が有する)

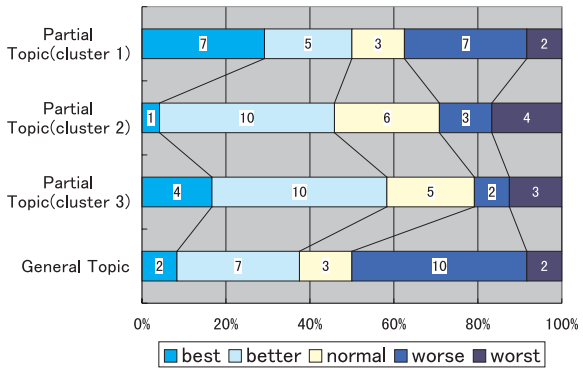


Fig. 3: Evaluation of topics.

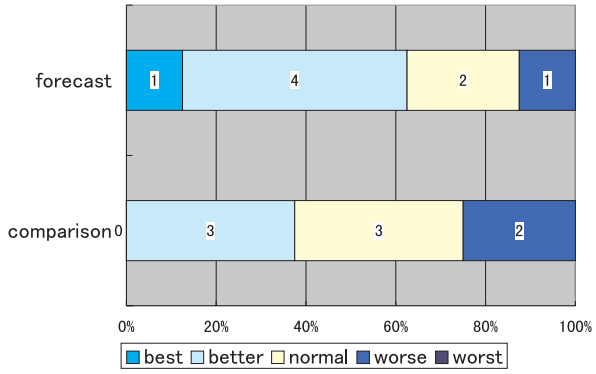


Fig. 4: Evaluation of summary.

学を専攻する大学院生 8 名である．実験データは，28 テキストを 8 クラスタに分類した内の 3 クラスタに含まれる 6 テキストである．評価は 5 段階評価で行なう．

6.1 話題の評価

部分話題 (Partial Topic) と全体話題 (General Topic) について評価した．評価対象は部分話題として 3 クラスタから 3 話題，全体話題として 3 話題，計 12 話題である．結果を Fig. 3に示す．表中の数字は，3 話題の評価の合計である．

部分話題では肯定評価を得た．各クラスタの部分話題に対し，評価者の半数が肯定評価 (‘best’,‘better’) を示した．肯定評価の割合は，最大 58.3%(cluster 3)，最小 45.8%(cluster 2) である．また，肯定評価に対する否定評価 (‘worse’,‘worst’) の比率 (否定評価数/肯定評価数) を Table 3に示す．各部分話題とも比率は低いことから，提案手法は部分話題の特定に有効である．

一方，全体話題では否定評価となった．Table 3の肯定評価に対する否定評価の比率では，部分話題の値に比べ大きな値となった．この原因として，全体話題の対象とするテキスト数の影響がある．本研究では，テキスト集合全体の代表として全体話題を特定する．しかし，対象テキスト数が多い場合，全体話題がテキスト集合全体を代表しない場合がある．今後，全体話題にとって適切なテキスト数について検討する．

6.2 要約の評価

要約の評価として次の点について調査した．

内容予測 (forecast)
　要約からテキスト集合の内容を予想できるか．
比較 (comparison)
　要約から予想した内容と実際のテキストの内容は同じか．

内容予測では評価者の 62.5% が肯定評価を示した．一方，要約から予想した内容と実際のテキストとの比較では，肯定評価は 37.5%，否定評価は 25.0% となった．評価者の回答を Fig. 4に示す．

予想と実際の内容との比較で否定評価となった原因に，不適切な文脈の影響がある．実装システムでは焦点の選択をランダムとした．このため，不適切な文脈を持つ要約が生成されることがある．今後，一貫した文脈生成における，焦点の選択方法について検討する．

7 結論

概要把握のための複数テキスト要約手法を提案した．大量のテキスト集合からの情報収集では，テキスト集合の概要把握が重要である．本稿では，話題特定と要約生成による概要把握を目的とした複数テキスト要約手法を提案した．話題と要約について評価実験を行なった結果，部分話題の特定，要約からの内容予測に提案手法の有効性を確認した．今後は，全体話題の対象とするテキスト数，焦点の選択基準について検討する．

参考文献

[1] 武田: ネットワークを利用した知的情報統合; 人工知能学会誌, Vol. 11, No. 5, pp.680-688(1997)
[2] 佐藤 円, 佐藤 理史, 篠田: 電子ニュースのダイジェスト自動生成; 情報処理学会論文誌, Vol.36, No.10, pp.2371-2379(1995)
[3] Hearst, M.A. and Pederson, J.O: Scatter/Gather as a tool for navigation of retrieval results; *1995 AAAI Fall Symposium on AI Application in Knowledge navigation*, pp.65-71(1995)
[4] 池田: 一変量の統計量; 統計ガイドブック, 新曜社, pp.50(1989)
[5] 福原, 武田, 西田: 統計情報と概念知識を用いたテキスト間の話題特定; 電子情報通信学会 信学技報 (AI98-62), Vol.98, No.498, pp.1-8(1999)
[6] 長尾, 佐藤, 黒橋, 角田: 自然言語処理; 岩波書店 (1996)