新聞記事を対象にした，検索，分類，複数文書要約システム ELIOTシステム

村上浩司† 野畑周‡ 関根聡§ 井佐原均‡
北海道大学† 通信総合研究所‡ ニューヨーク大学§

1 はじめに

新聞記事において情報を探す際には，キーワードを入力し，いわゆる情報検索により関連した文書を見つける方法が主流となっている．しかし，単なるキーワード検索だけでは非常に多くの検索結果が出力されてしまうことが少なくない．このような場合，検索された多くの文書の中からユーザが目的とする情報がどこにあるのかを求めることが非常に困難である．多くの情報をコンパクトに見るために複数文書要約の研究がされているが，検索された文書は数的に大量であるだけでなく，複数のサブトピックが混在しているため，そのまま複数文書要約を行うことが適さない状態にある．

提案するELIOTシステムは，こうした問題を解決するために，検索，テキスト分類，重要キーワード抽出と複数文書要約の技術を結合したシステムである．まず，ユーザがキーワードを入力し文書を検索すると，文書が大量にある場合システムは検索された文書をサブトピック毎に分類する．同時にそれぞれのサブトピックに対する重要キーワードが表示され，ユーザはそれらのキーワードからサブトピックが推測できるようになっている．そして興味のあるトピックを指定すると，そのトピックに属する複数の文書から1つの要約を作成し表示する．

このような検索，分類，要約の統合システムには，QCS[5]や，オンラインの新聞記事を対象としているNewsblaster[7]，NewsInEssence[10]などが知られている．ELIOTシステムでは，重要キーワードをデータから抽出しユーザに表示することでユーザの必要なサブトピックのみを対象に要約を行う．

2 構成

ELIOTシステムの構成を図1に示す．まず，ユーザによって入力されたキーワードが含まれる記事文書を検索する．

次に，検索によって得られた文書に対して形態素解析を行い，文書中の名詞を抽出すると同時に，それらの出現頻度を求める．これらの情報を用いて対象の文書群をクラスタリングにより分類する．この結果，各クラスタには類似した内容であると考えられる複数の文書が属する．

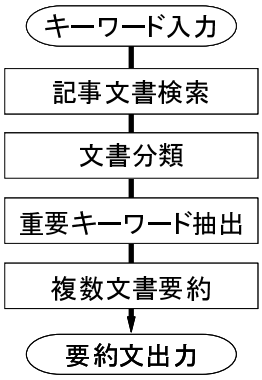

図1: ELIOTシステムの構成

そして各クラスタのサブトピックを示す代表的な重要キーワードをtf*idf値を用いて抽出する．各クラスタに属する記事文書のヘッドラインや重要キーワードから，ユーザが目的のクラスタを指定することで，そのクラスタに属する複数の文書から要約文が作成，出力される．

本システムは1998年，1999年の毎日新聞記事データを対象のコーパスとしている．

3 検索

ユーザは通常のキーワード検索と同様に，キーワードを入力する．システムは，入力されたキーワードを含む記事文書をコーパスから検索する．

検索された文書は入力されたキーワードを含むが，その文書の内容は目的の情報が記述されているものだけではなく，ユーザが必要としない情報である文書も数多く検索されることが多い．そのためユーザは大量の文書の多くを確認するか，もしくは更にキーワードを入力して検索を行わなければ，目的の文書に到達できない．また，検索により見つかる文書は数的に大量なだけでなく，複数のサブトピックが混在していることが多い．そこで，検索から得られるすべての文書から

ユーザが目的の文書を見つけるのではなく，類似したサブトピックの文書をクラスタリングにより分類することでユーザの目的文書選択のコストを削減する．

4 クラスタリングによる文書分類

新聞記事などの一般的な文書の分類においては，それぞれのサブトピックを特徴付けるのは名詞であることに着目する．そこで，検索された文書を JUMAN[2] により形態素解析を行い名詞を抽出する．同時に，対象文書全体と各文書における各名詞の出現頻度もそれぞれ求める．文書 u および v 間の類似度 $sim(v,u)$ は，以下のように cosine 関数として定義し，先に求めた各文書中の名詞の頻度から計算する．

$$sim(v,u) = \frac{\sum_{i=1}^{T} v_i \cdot u_i}{\sqrt{\sum_{i=1}^{T} v_i^2 \times \sum_{i=1}^{T} u_i^2}}$$

我々はクラスタリングのアルゴリズムとして k-way クラスタリング法を用いる．この手法では，まず入力されたデータを 2 つのクラスタに分類する．そして分割されたクラスタを再び 2 つのクラスタに分割していく．こうした処理を必要なクラスタ数になるまで繰り返す．これらの分割処理は 2-way クラスタリングが，以下に示す識別関数 [3, 4] を最適化するように行われる．

k はクラスタ数を，S_i は i 番目のクラスタに属する文書集合を表す．

$$maximize \sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_i} sim(v,u)}$$

本システムは，以上のようなクラスタリングを実現するために，ミネソタ大学で開発されているクラスタリングツールキット CLUTO [1] を用いた．

5 重要キーワード抽出

ユーザが，生成されたクラスタから目的の情報を選別できるように，各クラスタの特徴を表すキーワードを抽出し，ユーザに提示する必要がある．ここで抽出されるキーワードは，各クラスタの内容を適切に反映し，他のクラスタと識別できる名詞である必要がある．そのため文書中に含まれる名詞が，低頻度で複数のクラスタに分散して出現すれば非重要名詞であるが，高頻度で特定のクラスタに集中して出現すれば，そのクラスタを特徴付けることになる．そこで各クラスタに属する名詞の tf 値と idf 値をクラスタ内とクラスタ間で独立して取り扱う．

各クラスタに属する名詞のクラスタ間における名詞の it*idf 値 (gTF·gIDF)，クラスタ内の名詞の tf 値 (cTF) および，各名詞が特定のクラスタに出現する割合，つまり出現密度をカバレッジ $cov(= cTF/gTF)$ として導入し，これらの値を用いて重要キーワードを抽出する．

i 番目のクラスタにおける j 番目の名詞 NN_{ij} に対する重要キーワード判定を図 2 に示す．$\sigma, \delta, \theta, \phi, \rho$ はそれぞれ閾値を示す．

begin
$cov_{ij} = cTF_{ij}/gTF_{ij}$;
if$((cTF_{ij} >= \sigma)$ && $(gTF_{ij} >= \delta)$ &&
$(gTF_{ij} * gIDF_{ij} >= \theta)$ &&
$((gIDF_{ij} >= \phi)$ || $(cov_{ij} >= \rho)))$ **then**
名詞 NN_{ij} を重要キーワードとして決定;
end

図 2: 重要キーワードの判定

6 複数文書要約

本システムの要約部では，重要文抽出に基づく複数文書要約システムを用いている．この要約システムは，2002 年に行われた自動要約システムの評価型ワークショップ The Second Text Summarization Challenge (TSC-2)[8] に参加したものを用いている [9]．重要文抽出は要約の要素技術の一つであり，文章中の各文について重要度を見積り，その値に従って要約に必要な情報を含む重要な文を抜き出すものである [11, 6]．

この要約システムでは，各文の重要度を記事中の文の位置・文長・文中の単語の頻度・見出しとの関連度の 4 つの関数を用いて求める．これら各関数 $Score_j$ の値に重み付け α_j を与え，それらの和が各文 S_i の重要度となる:

$$\text{Total-Score}(S_i) = \sum_j \alpha_j Score_j(S_i)$$

重みの値は，動的に生成された記事セットに対して事前に学習することはできないので，TSC-2 で用いたものをそのまま用いている．

6.1 文の位置

文の位置情報に基づく関数は，文の位置の逆数を与えるものである．つまり i 番目の文 S_i に対するスコアは，

$$Score_{\text{pst}}(S_i) = \frac{1}{i}$$

となる．これは記事の先頭に近い文ほど重要度が高いことが多い，という観測事実に基づいて定義されたものである．

6.2 文の長さ

各文の長さに基づく関数は，長さ L_i が一定の値 C より短い文にはペナルティとして負の値を与えるものである．文の長さは文字数で表している：

$$\begin{aligned} Score_{\mathrm{len}}(S_i) &= 0 \quad (L_i \geq C \quad \text{のとき}) \\ &= L_i - C \quad (\text{それ以外}) \end{aligned}$$

このペナルティは，極端に短い文は重要文として選択されることが非常に稀であるという観測事実に基いている．ペナルティを与えるしきい値 C は，TSC データを用いた実験の結果からを 20(文字) とした．

6.3 tf*idf 値

文中の単語の頻度を用いる関数は，クラスタリングの重要キーワード抽出と同様に，各文中の単語について tf*idf 値を計算し文のスコア付けを行うものである．tf*idf 値は，各記事中の単語 w の頻度 $tf(w)$ と，その単語がある記事群の中で現れた記事の数，すなわち記事頻度 $df(w)$ とを組み合わせて計算される値で，記事中のある単語がどの程度その記事特有の単語であるかを示す．記事数 DN 個の記事群が与えられたときの，各単語における tf*idf 値の計算式は以下のようになる：

$$\mathrm{tf*idf}(w) = tf(w) \log \frac{DN}{df(w)}$$

各文のスコアは，文中の各単語に対する tf*idf 値の和によって与えられる：

$$Score_{\mathrm{tf*idf}}(S_i) = \sum_{w \in S_i} \mathrm{tf*idf}(w)$$

6.4 見出し

記事の見出しを用いる関数は，各記事の見出しに含まれる単語に対する tf*idf 値を用いて文のスコア付けを行うものである．これは「見出しと類似している文は重要である」という仮定に基いている．文 S_i 中の対象単語について，その名詞が見出し H に含まれていれば，その tf*idf 値を文のスコアに加算する．文のスコアを与える式を以下に示す：

$$Score_{\mathrm{hl}}(S_i) = \frac{\sum_{w \in H \cap S_i} \mathrm{tf*idf}(w)}{\sum_{w \in H} \mathrm{tf*idf}(w)}$$

6.5 類似した文の除去

複数の文書を要約する際には，類似した文を除き，生成される要約に冗長な部分が含まれないようにすることが必要である．本要約システムでは，文書セット中の文の各組み合わせについて Dice の係数に基づき類似度を求めておき，ある文が重要文として選択されたときにその文に対し一定のしきい値以上の類似度をもつ文は，重要度に係らず要約文の出力から除かれるようにしている．

図 3: ELIOT メイン画面

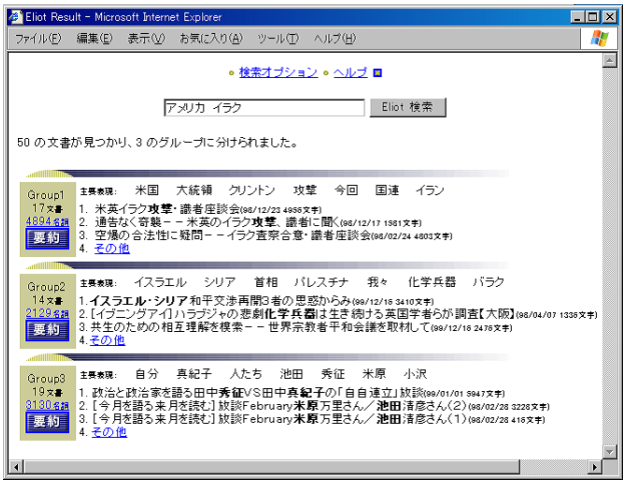

図 4: 分類結果

7 画面サンプル

実際のシステムの動作例を示す．図 3 に ELIOT システムのメイン画面を示す．キーワード入力ができるほか，クラスタ数をユーザによって指定，もしくは自動分類を選択できる．ここでは “アメリカ” および “イラク” を入力キーワードとした．

入力キーワードによって文書を検索し，対象となる文書をクラスタリングにより分類した結果を図 4 に示

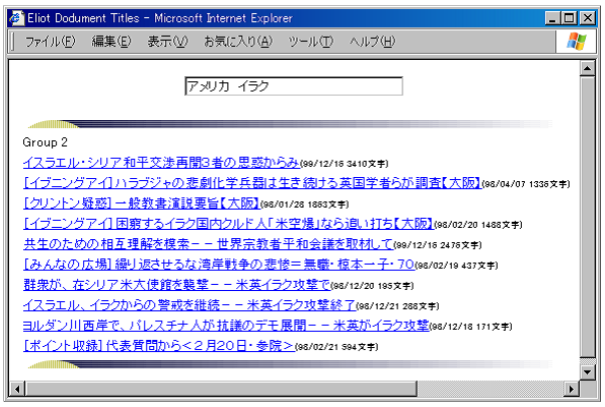

図 5: クラスタに属する記事一覧

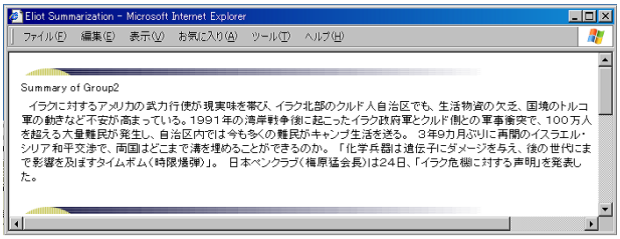

図 6: 要約結果

す．この例においては，クラスタ数は自動で決定した．ここでは 50 の文書が見つかり，“アメリカ，攻撃，国連” に関するクラスタ，“イスラエル，シリア，パレスチナ” に関するクラスタ，“評論家による放談” に関するクラスタの 3 つに分類された．各クラスタについて，属する文書数と名詞数，クラスタの重要キーワード，および重要キーワードを多く含む文書のタイトルが表示される．

図 4 の分類結果から，指定した Group2 に属する全ての記事文書ヘッドラインを図 5 示す．各記事の内容を参照する場合はこのヘッドラインから見ることができる．

図 6 に，図 4 で Group2 を指定して複数文要約を実行した結果を示す．

8 まとめ

キーワードから新聞記事を検索すると，類似した話題の記事が多く得られることがある．そしてユーザが必要とする情報もその中の 1 つの話題に存在していることが多い．そのため記事検索を行うと同時に，クラスタリングにより類似する話題の記事を自動的に分類し，数多く見つかる文書そのものではなく，少数のクラスタからユーザが選択することで，より効率良く必要な情報を収集できると考えられる．そこで筆者らの複数文自動要約システムにこのクラスタリングを適用したシステムを開発した．

参考文献

[1] *http://www-users.cs.umn.edu/ karypis/cluto/.*

[2] *JUMAN, http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html.*

[3] *Criterion Functions for Document Clustering: Experiments and Analysis*, Minneapolis, MN, 2001.

[4] *Evaluation of Hierarchical Clustering Algorithms for Document Datasets*, Minneapolis, MN, 2002.

[5] Daniel M. Dunlavy, John Conroy, and Dianne P. O'leary. Qcs: A tool for querying, clustering, and summarizing documents. In *Human Language Technology Confernece of the North American Chapter of the Association for Computational Linguistics(HLT-NAACL) 2003 Demonstrations*, pp. 11–12, May-June 2003.

[6] I. Mani and M. Maybury. *Advances in Automatic Text Summarization*. The MIT Press, Cambridge, MA, 1999.

[7] Kathleen McKeown, Regina Barzilay, John Chen, David Elson, David Evans, Judith Klavans, Ani Nenkova, Barry Schiffman, and Sergey Sigelman. Columbia's newsblaster: New features and future directions. In *Human Language Technology Confernece of the North American Chapter of the Association for Computational Linguistics(HLT-NAACL) 2003 Demonstrations*, pp. 15–16, May-June 2003.

[8] NII, editor. *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, Tokyo, Japan, October 2002. National Institute of Informatics.

[9] C. Nobata, S. Sekine, K. Uchimoto, and H. Isahara. A summarization system with categorization of document sets. In *Working Notes of the Third NTCIR Workshop Meeting*, pp. 33–38, October 2002.

[10] Dragomir R. Radev, Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *Human Language Technology Conference*, San Diego, CA 2001.

[11] 奥村学, 難波英嗣. テキスト自動要約に関する研究動向 (巻頭言に代えて). 自然言語処理, Vol. 6, No. 6, pp. 1–26, 1999.