テキストマイニングによる機器異常診断支援の試み

吉田　稔† 中川　裕志† 渋谷　久恵†† 前田　俊二††

† 東京大学情報基盤センター　〒 113–0033 東京都文京区本郷 7-3-1
†† 日立製作所横浜研究所　〒 244-0817 神奈川県横浜市戸塚区吉田町 292
E-mail: †mino@r.dl.itc.u-tokyo.ac.jp, nakagawa@dl.itc.u-tokyo.ac.jp,
††{hisae.shibuya.mp, shunji.maeda.ap}@hitachi.com

あらまし　設備機器運用の際に電子的に蓄積された管理レポートを、実際の機器運用に役立てるための、テキストマイニング技術を応用した支援システムの構築について報告する。提案システムは、文書集合をトピックモデルにより分析し、各トピックの要約を提示することで、文書全体を俯瞰することを補助する。
キーワード　テキストマイニング、トピックモデル、文書要約

A Text Mining Approach to Diagnosis of Instrument Anomalies

Minoru YOSHIDA†, Hiroshi NAKAGAWA†, Hisae SHIBUYA††, and Shunji MAEDA††

† Information Technology Center, University of Tokyo
7–3–1 Hongo, Bunkyo, Tokyo, 113–0033 Japan
†† Yokohama Research Laboratory, Hitachi, Ltd.
292, Yoshida-cho, Totsuka-ku, Yokohama-shi, Kanagawa, 244-0817 Japan
E-mail: †mino@r.dl.itc.u-tokyo.ac.jp, nakagawa@dl.itc.u-tokyo.ac.jp,
††{hisae.shibuya.mp, shunji.maeda.ap}@hitachi.com

Abstract　We propose a diagnosis system for instrument anomalies that can suggest possible actions for future troubles by automatically analyzing previous reports on maintenance of the instruments. Our system uses a topic model to cluster words and documents in the text, and presents summaries of each topic by using word frequencies and summarized documents.
Key words　Text Mining, Topic Models, Document Summarization

1. はじめに

本稿では、「機器メンテナンスに関するテキストデータ」をマイニング対象としたテキストマイニングシステムを報告する。本研究で対象とする機器には、各種センサーにより検知された値に基づき異常を知らせる警報が装備されており、この警報が発せられた場合、担当者が、当該機器のメンテナンスを行う。マイニング対象とするのは、この機器のメンテナンスの際に蓄積される業務レポートテキストデータである。

各テキストデータには、「具体的にどのような状況であったか」、また、「それに対してどのように対処を行ったか」がテキストで記述されているため、これらを自動的にマイニングすることで、「過去にどのような事象が起こり、それに対してどのような対策を講じたか」を俯瞰することが可能となり、将来のメンテナンスをより効率的に行えることが期待できる。

レポートは、主に、様々な**事象**（「どんな種類の警報が鳴っているのか」「どんな原因が推測されるか」「どんな種類の対処法を取ったか」等）が混合されてできており、これらが組み合わさって一つの**シナリオ**となる。（「警報 A が鳴る」→「原因 B を推測」→「対処法 C を適用」、等。）本研究では、「文書集合中にどのような事象が記述されているか」という「事象のマイニング」と、「実際のイベントで、それらの事象がどのように組み合わさっているのか」という「シナリオマイニング」の 2 つの側面からのテキストマイニングを行うシステムを提案する。

提案システムは、近年普及の著しい**トピックモデル**と、自然言語処理分野で盛んに研究される**文書要約技術**を組み合わせることで、効率的な文書集合の俯瞰を可能とする。トピックモデルとしては、近年広く用いられる Latent Dirichlet Allocation (LDA) [3] を，文書クラスタリングに応用できるよう拡張した **Dirichlet-Enhanced Latent Semantic Analysis** [16] (以下，DELSA) と呼ばれるモデルを用いる．DELSA を用いることで、それぞれの**事象を「単語トピック」**として、また、それぞれの**シナリオを「文書トピック」**として推測できる。

より具体的には、提案システムは、

- DELSA により、各文書の文書トピック（クラスタ）、及び、各単語の単語トピック（クラスタ）を同時に推定する。
- 各トピックを、文書要約技術により俯瞰する

という２段階の処理を行うシステムである。

近年のトピックモデルの普及に伴い、これを要約に応用する研究が盛んとなっている [4] [10] [9] [12]。既存研究は、多くの場合、複数文書要約を対象としている。通常、複数文書要約 (Multi Document Summarization, MDS) は、「**同一事象を記述した複数文書**から、要約を生成する」という問題設定である。提案システムでも、複数文書を要約することが目的であるが、入力として与えられるレポート集合は、雑多な話題が混在した集合であり、要約すべき「同一の話題の文書集合」は与えられておらず、このため、「文書集合を自動的に発見し、その結果を自動的に要約する」ことが必要である。また、多くの MDS の設定では、文書は文の集合に分割され、その中から選ばれた文を要約に用いるが、本研究では、それぞれの文書は別々の事象を表現しているため、「**各レポートを、文単位で分解することができない**」という制約が必要となる。何故なら、例えば、もしも「事象 A」と「対処法 B」を別々のレポートから取得し合成した場合、実際にはあり得ないような事象・対処法の組み合わせを提示してしまい、却って人間の判断の妨げになってしまう危険性が高いためである。従って、本研究における要約は、

（1）「類似する**文書**のクラスタ」を形成する。

（2） 各クラスタから代表的な文書を抜き出し、各文書を要約する。

という処理となる。

文書要約においては、「文選択」(extraction) と「文圧縮」(compression) という２つの要約方法があり、また、これら２つを組み合わせた手法も存在する [8]。本研究でも、文選択・文圧縮を共に行うが、文選択については、文書要約において広く用いられる MMR [13] を用い、文圧縮においては、木構造マイニングを利用した手法を用いる。

2. 問 題 設 定

入力として、文書コレクション $\mathcal{D} =< d_1, d_2, \cdots, d_n >$ が与えられるとする。このとき、d_i には、それに付随する木構造 t_i が与えられている。各 t_i は、d_i 中の各文節の依存構造グラフであり、文節間の依存構造と、文節および、それを単語分割及び形態素解析した結果の単語リストを保持する。例えば、d_i が「問題発生した機器について復帰完了した」という一文であった場合、その依存構造木(注1)は、(“復帰完了した”, (“機器について”, (“問題発生した”))) となり、各ノード（文節）にはさらに、それを形態素解析した結果の単語リストが付随する。

2.1 使用データ

本研究で使用するデータセットは、表形式で提供されており、テキスト本文のほかに、警報の種類、警報の起こった日付等のデータが付与されている。テキスト本文には、状況および対応の様子が記述されており、多くの場合は時系列に沿って文が並んでいる。また、本文に加え、「アラーム」や「対策」といった列があり、それぞれ、どの警報が鳴ったのか、および、どのような対策が行われたのかがまとめられている。提案システムでは、CaboCha [11] により、各文に対し単語分割・形態素解析および依存構造解析を行う。本研究では、形態素解析の結果を利用し、各文から名詞のみを取り出し(注2)、後述のトピックモデル及び木構造マイニングの対象とする。

以下、具体的な手法について解説する。手法は、「STEP-1: DELSA によるトピック推定」と、「STEP-2: 各トピックの要約」の２段階に分かれ、STEP-1 の結果としてギブスサンプリングによって得られる各単語の単語トピック番号、および各文書の文書トピック番号が出力され、STEP-2 の入力として与えられる。

3. トピックモデルによる単語・文書トピックの推定

3.1 Latent Dirichlet Allocation (LDA)

LDA は，文書等の，スパースなベクトルを効率良くモデル化するための生成的確率モデルである．LDA においては，文書中の各単語は，**トピック**から生成されると仮定される．ここでトピックとは，単語の出現確率を表す分布（多項分布）であり，例えば，電気に関するトピックは「電圧」や「kV」といった単語に高い確率を与える分布，対策一般に関するトピックは「推定」や「対策」といった単語に高い確率を与える分布として表現される．LDA を用いることにより，文書中の各単語に対し，その単語がどのトピックから生成されたかの推定値（トピックの番号）を与えることができる．

3.2 Dirichlet-Enhanced Latent Semantic Analysis (DELSA)

LDA においては，各文書に一つずつ**トピック分布**（＝各トピックの出現確率を表す多項分布）が与えられるが，このトピック分布は，単一の**ディリクレ分布**から，それぞれ独立に生成される．DELSA では，このディリクレ分布の代わりに，**ディリクレ分布を基底として持つディリクレ過程**[6] を用いて，各文書のトピック分布を生成する．ディリクレ過程は，離散確率分布を生成する確率分布（確率分布の確率分布）であり，基底分布と呼ばれる分布（この場合，ディリクレ分布）から，可算無限個のサンプル（この場合，ディリクレ分布からのサンプル，すなわち，多項分布）と，その混合比を生成することができる(注3)．すなわち，この場合，ディリクレ分布から可算無限個のトピック分布が，その混合比とともに生成されることになる．その後，混合比にしたがって，各文書のトピック分布が選ばれる．

DELSA における文書の生成過程を，以下に示す．

（1） ディリクレ過程から，可算無限個のトピック分布と，

(注1)：本稿では、各木構造は S 式で表現する。

(注2)：本研究で対象としている機器異常診断レポートにおいては、「見る」「食べる」等の一般的な動詞の頻度は少なく、「実施する」「調査する」等の、「サ変名詞＋する」形式の表現が多かった。このため、名詞だけを取得すればマイニングに十分なデータが得られると判断した。

(注3)：このとき，混合比（各サンプルの確率分布）は，基底分布と似た分布となる．

その混合比を生成.

（2） 各文書毎に，1. で得られた混合トピック分布（＝混合多項分布, Multinomial Mixture）の中から，トピック分布を一つ選ぶ.

（3） 文書内の各単語のトピックを，得られたトピック分布（＝多項分布）からサンプルする.

（4） 各単語を，その単語に割り当てられたトピック（＝単語比率を表す多項分布）から生成する.

上記第 2 ステップで，各文書毎にトピック分布が一つ選ばれるが，このとき，同一のトピック分布が選ばれた文書は，同一の話題に属すると考えることができる．すなわち，同一のトピック分布が割り当てられた文書同士を同一のクラスタにまとめることで，文書クラスタリングを行うことができる．このように，DELSA では,「各単語に割り当てられるトピック（＝単語分布）を推定することによる単語クラスタリング」と「各文書に割り当てられるトピック分布を推定することによる文書クラスタリング」を同時に行うことができる.

事前分布としてディリクレ過程を用いる利点としては,

- ディリクレ過程は,「同一の点が複数回選ばれる確率が高い」という性質があるため [2]，文書がまとまり易い.
- 可算無限個の点を生成するため，文書クラスタの数に上限を設定する必要がない.

というものがあげられる．特に後者は，文書集合を大きくしても，文書クラスタ数を人手で調整する必要がないという点で，有利である.

以下，具体的な生成過程を示す．各文書 d 中の単語 $w_{d,n}$ を，その単語に割り当てられたトピック $z_{d,n}$ から以下のように生成する(注4).

$$w_{d,n} \sim Multi(\phi_{z_{d,n}})$$
$$z_{d,n} \sim Multi(\theta_d)$$

ここで，パラメータ ϕ_i は，i 番目のトピックにおける単語分布を表すベクトルであり，$\sum_j \phi_{i,j} = 1$ である．各 ϕ_i は，単一のディリクレ分布 $Dir(\beta)$ から生成される.（パラメータ β は，$|V|$ 次元ベクトル．ここで $|V|$ は単語の種類数.）また，θ_d は，文書 d のトピック分布を表すベクトルであり，$\sum_i \theta_{d,i} = 1$ である.

ここで，LDA では，θ_d がディリクレ分布から生成されていたが，DELSA では，以下のような（ディリクレ過程から生成された）分布 G から生成される.（混合トピック分布から，1 つのトピック分布を選ぶ.）(注5)

$$\theta_d \sim G$$
$$G \sim DP(G_0, \alpha)$$

以上のモデルに基づき，各単語のトピックを推定する．トピックの推定には Collapsed Gibbs Sampling [7] を用いた．ここで，DELSA においては，通常の LDA での単語トピックラベル $z_{d,n}$ に加え，文書のトピック分布ラベルを表す変数 c_d を用意し，サンプリングする必要があることに注意されたい．得られた c_d を利用し，同一の c_d を持つ文書を一つのクラスタにまとめる.

LDA をディリクレ過程で拡張したモデルとしては，Hierarchcal Dirichlet Process (HDP) [15] があるが，HDP が，**トピック**を可算無限個生成し，トピックを文書間で共有するモデルなのに対し，DELSA では，トピック数は有限のまま，文書の**トピック分布**を可算無限個生成し，トピック分布を文書間で共有する．トピック分布は各文書に一つのみ与えられるため,「同じトピック分布を共有する文書どうしをまとめる」ことで，自然に文書クラスタリングが行えるという利点がある.

4. 要約によるトピック俯瞰

DELSA は、単語トピック、文書トピックを同時に推定するモデルである。これは、本研究で提案するシステムにおいて、単語の側面と、文書の側面という 2 つの視点から文書集合を要約できることを意味する。本節では、実際にこれらのトピックを要約表示する手法について説明する。

4.1 単語クラスタ（トピック）の要約

通常、トピックモデルにおいて、得られた単語トピックの性質を俯瞰するためには、各トピックにおける確率上位の単語を列挙することが広く行われている。これに類似した結果が得られる手法として、「ギブスサンプリングにより各単語にトピックラベルを付け、それを集計し、各トピックの頻出単語を列挙する」という手法も取ることができる。本研究では、後者の手法を用い、各単語トピック毎に頻出単語リストを表示することで、単語トピックの要約とする。

4.2 文書クラスタ（トピック）の要約

文書クラスタにおいては、各クラスタが「単語の分布」ではなく「トピックの分布」で表現されているため、「確率の高いトピックを列挙する」という方法では、間接的な表現方法となり、直感的にわかりづらい表現となってしまう。

また、単語トピックと同様に、「ギブスサンプリングにより各文書にトピックラベルを付け、それを集計し、各トピックの頻出文書を列挙する」という手法も考えられるが、これも、そのままでは文書トピックに適用することは困難である。何故なら、文書集合中で「全く同一の文書」が出現することは稀であり、同一の文書が出現しなければ、各文書の頻度はすべて 1 となり、集計しても文書トピックを特徴づけることが困難であるからである。

このため、本研究では、トピックを**文書要約**により俯瞰する手法を提案する。これは、各文書トピックにおいて、そのトピックラベルを付与された文書を集めてクラスタを作り、そのクラスタに対し、文書要約の技術を適用することで、クラスタ、ひいてはトピックを代表するような文字列を得ることを目的とする。

本研究では、クラスタから、それをできるだけ代表するような文書を数個選び、それをクラスタの要約として表示する「選

(注4)：$Multi$ は多項分布を示す.

(注5)：G_0 は基底分布，α は集中度パラメータと呼ばれる，クラスタの集中度を制御できるパラメータ.

択による要約」と、各文書をより簡潔に表現する「圧縮による要約」の 2 種類の要約を考える。

4.2.1 文書選択による要約

文書選択による要約には、文書要約において広く用いられる手法である、MMR [13] を使用する。MMR は、以下の式で定義されるスコアである。

$$MMR(x, A) = \lambda sim(x, q) - (1 - \lambda)\max_{y \in A} sim(x, y)$$

この MMR スコアが高い文書を順に選択していく。ここで、x は評価対象の文書、A は既に選択された要約の集合である。ここで q はクエリ（何らかの理想とする文書）であるが、本研究では、これをクラスタの重心とする。MMR の第一項は「文書の適合度」を、第二項で「すでに選択された文書との類似性」を表す。第一項の値が大きく、第二項の値が小さいほど、「適合度が高く、しかもいままで選ばれていない」内容の文書であり、要約の候補として適切ということになる。すなわち、よりクラスタの重心に近い文書で、さらにすでに選択された文書と違いの大きい文書が、新たな要約として選択されることになる。また、λ は、第一項と第二項の重みを調整するパラメータである。

4.2.2 文圧縮による要約

a） 部分木の抽出

本ステップでは、木構造マイニングを文書集合 D に適用することにより、頻出木構造を取得する。ここで頻出木構造とは、D 中の依存構造木の集合で、しきい値（現在は 2）以上の回数出現する部分木構造のことである。(注6)

問題設定の項で述べたとおり、各文には依存構造木 t_i が付随している。依存構造木の各ノードは、文節を表しており、エッジは文節どうしの依存関係を表す。例えば、「故障発生した機器について調査完了した」という文の依存構造解析結果は、([調査, 完了], ([機器], ([故障, 発生]))) となる。また、同一クラスタ内に、別の文、例えば「問題発生した機器について復帰完了した」という文が存在した場合、その構造は ([復帰, 完了], ([機器], ([問題, 発生]))) となる。クラスタがこの 2 つの文で形成されていた場合、例えば部分木として ([完了], ([機器], ([発生]))) が抽出でき、その頻度は 2 となる。このようにして得られた頻出木を部分木として持つ文を発見し、そこから要約を生成する。

さらに、文書内の複数文を一つの木で表すことにより、頻度の高い**イベント連鎖**を取得できるようになる。例えば、「警報 A が発生」「装置 B を修復」という表現が共に出現する文書が多かった場合、「((発生, (警報 A)) (修復, (装置 B)))」という木構造の頻度が高い、という事象としてマイニングできることになる。頻出木構造の取得には、FREQT [1] アルゴリズムを用いる。

b） 頻出部分木による要約生成

頻出木構造に基づき、D 中の依存構造木における各文節をスコア付けする。現在のシステムでは、各頻出木構造 t に、以下のスコアを付与する。

$$score(t) = |t| \cdot \log(freq(t))$$

ここで $|t|$ は、木構造 t のノード数を表す。各文節 c のスコアは、c を含む木構造ののスコアのうち、最大のものを、その文節のスコアとして定義する。

この文節のスコアに基づき、システムは、各構文木の木構造の根から、与えられた文長を超えないように、再帰的に子ノードを探索していく。各ノードは各文節に対応しており、ノードの探索は、スコアの高い文節に対応するノードから優先的に行われる。与えられた文長を超えた時点で探索は打ち切られ、それまで辿られたノードから文を生成する。

依存構造木からの要約文の生成は、日本語の依存構造が、「前の文節から後の文節にかかる」および「依存関係が交差しない」という性質を持つことから、自明に行うことができる。

5. 評　　価

生成された要約の評価には、**ROUGE スコア** [13] を用いる。ROUGE スコアとは、何らかの正解（**参照要約**と呼ばれる）を準備し、その要約の再現率をもとに機械生成要約の良さを評価する手法であり、ROUGE-N スコアは、以下で定義される。

$$\frac{\sum_{x \in s_r} \min(freq(x, s_m), freq(x, s_r))}{\sum_{x \in s_r} freq(x, s_r)}$$

ここで、s_m は、機械生成による要約、 s_r は参照要約、$x \in s$ は、x が s の部分として含まれる N グラム（N 単語連接）であることを示し、$freq(x, s)$ は、x の s 中に含まれる頻度を表す。

本研究では、レポート中の表形式における「アラーム」列および「対策」列を擬似的な参照要約として用いる。これらの列には、どのような問題が起こったか（どのようなアラームが反応したか）および、起こった問題に対してどのような対策を採ったかが簡潔に記述されており、本文中の問題および対策記述部分が、より簡潔な形で示されていると考えられる。我々は、各レポートについて、この「アラーム」列と「対策」列を参照要約とし、クラスタ上位 k 個のテキストが、クラスタ中の各レポートにおける参照要約をどれほど再現できたかを、レポート毎に ROUGE スコアで計算する。その後、これらの ROUGE スコアを平均することで、システム全体のスコアとする。ROUGE スコアは、再現率に基づくスコアのため、選ばれたテキスト中に、問題および対策に関連しない部分が含まれていても、スコアには影響しない。

文圧縮の際は、文字数の上限値 L を設定し、文の長さが L を超えるまで、ルートノードから開始し、子のノードに相当する文節を順次選択していく。このさい、文節にスコアが与えられている場合、最もスコアの高い文節を選択し、そうでない場合はランダムに文節を選択する。実験では、$L = 50$ とし、また、MMR のパラメータ λ は、$\lambda = 0.5$ と設定した。

単語および文書のトピックラベルは、ギブスサンプリングを実行(注7)し、その時点で各単語・文書に付与されたラベルを用

(注6)：ただし、レポートの中には、全く同じ長文が重複して登場することがあり、このとき、頻出木構造の種類が非常に多くなり、スコア付けに悪影響を及ぼすことがあるため、木構造のサイズに上限（現在は 7）を設けている。

(注7)：iteration 回数は 500 とした。

表 1　10 回試行による平均の ROUGE-1 スコア（単位：%）。Top-k は上位 k 件のテキストを使用したことを示す。TREE は木構造マイニングによる文圧縮、RANDOM は乱数による文圧縮を示す。

Algorithm	Top-1	Top-2	Top-3	Top-4	Top-5
文書選択：MMR					
TREE	32.81	43.28	48.84	52.29	54.75
RANDOM	32.35	42.51	47.91	51.67	54.08
文書選択：ランダム					
TREE	20.53	32.40	39.70	45.32	49.10
RANDOM	20.09	31.79	38.35	43.44	47.21
文書選択：オラクル（理想値）					
TREE	43.36	55.15	61.51	65.73	68.75
RANDOM	42.60	54.55	61.00	65.27	68.38

表 2　文長 $L = 25$ の場合。他の設定は表 1 と同様。

Algorithm	Top-1	Top-2	Top-3	Top-4	Top-5
文書選択：MMR					
TREE	24.53	34.22	39.13	42.67	45.38
RANDOM	21.66	31.23	35.87	39.84	42.48
文書選択：ランダム					
TREE	12.81	21.80	29.43	33.76	37.90
RANDOM	10.76	18.71	24.45	29.13	32.70
文書選択：オラクル（理想値）					
TREE	36.13	48.14	54.72	59.02	62.03
RANDOM	34.88	46.95	53.47	57.71	60.73

いる。(注8)

文圧縮の精度を調べるため、提案手法（部分木の抽出による手法, 表中 “TREE”）と、ランダムに文節を選択する手法 (表中 “RANDOM”) を比較する。また、文書選択手法には MMR を用いるが、比較対象として、ランダムに文書を選択する手法 (表中「ランダム」) と、参照要約を見て最も ROUGE スコアを上昇させる文書を選ぶ手法（表中「オラクル」）を用いた。

表 1 に結果を示す。MMR による代表文書抽出は、Top-1 ではランダムとオラクルの平均程度の値を示したが、Top-5 まで文書を増やすにつれ、オラクルとの差が拡大している。このため、複数文書の選択において、MMR よりもより効率的な手法を用いる必要があると考えられる。文圧縮による要約抽出は、ランダムに文節を選択する手法よりは高いスコアとなったが、その差はそれほど大きくならなかった。この結果は、現在の文節選択手法にはまだ工夫の余地があることを示唆していると考えられる。また、原因の一つとして、要約文の文長 L の長さが十分大きく、ランダムに文節を選択しても適切な文節を選択してしまうという可能性も考えられる。そこで、文長 L を 25 とした場合のスコアを同様に計測した（表 2）。$L = 25$ においては、“TREE” と “RANDOM” の差は $L = 50$ の場合よりも大きくなった。これにより、要約文が十分に短い場合は、ある程度、木構造による文圧縮の効果が見られることが観察された。

また、DELSA によるクラスタリングの傾向を調べるため、トピックを用いず、単語の出現確率で直接文書をモデル化する Dirichlet Process Unigram Mixture (DPUM) [14] で同様のクラスタリングを行ったところ、DPUM のほうがクラスタ数が少なくなる（DELSA の平均クラスタ数 73.4 に対し、DPUM の平均クラスタ数 24.2 個）傾向が見られた。単語をそのままクラスタリングに用いた場合、特に今回用いた文書集合では、アラーム名が曖昧性の少ない形で高い頻度で記述されているため、同じアラーム名に対応する文書がまとまり易い傾向が見られた。これに対し、DELSA を用いたクラスタリングでは、直接単語を用いないため、この影響が緩和され、同じアラーム名に対応する文書の中で、異なる内容のもの（異なる現象や、異なる対策が記述されたもの）が別々のクラスタにまとめられている様子が観察された。

(注8)：予備実験においてギブスサンプリングと変分ベイズ法 [16] の両方を試した結果、前者のほうが良い結果を示したため、本研究では前者を用いる。

6. おわりに

トピックモデルと、その要約表示を用いた、機器異常診断レポートに対するテキストマイニングシステムの提案を行った。トピックモデルとして、文書トピックと単語トピックを同時に推定できるモデル DELSA を利用し、さらに、文書トピックを要約表示するシステムを構築した。今後は、より広範囲な分野のテキストを対象とした適用性の検証や、日付データなどのメタデータとテキストマイニング結果の融合等について検討を行なっていく予定である。

文　献

[1] Kenji Abe, Shinji Kawasoe, Tatsuya Asai, Hiroki Arimura, Setsuo Arikawa, Optimized Substructure Discovery for Semi-structured Data, Proc. 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2002), LNAI 2431, pp. 1–14, (2002).

[2] Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. Annals of Statistics, vol. 2, no. 6, pp. 1152–1174, (1974)

[3] D.M.Blei, A.Y.Ng, M.I.Jordan, ”Latent Dirichlet Allocation” JMLR, vol.3, pp.993-1022 (2003)

[4] Asli Celikyilmaz, Dilek Hakkani-Tur. Discovery of Topically Coherent Sentences for Extractive Summarization. ACL 2011, pp. 491–499 (2011)

[5] Douglass R. Cutting and David R. Karger and Jan O. Pedersen and John W. Tukey, Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, SIGIR’92, pp. 318–329, (1992)

[6] Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics, Vol. 1, No. 2, pp. 209-230, (1973)

[7] Griffiths, T. L., Steyvers, M. A probabilistic approach to semantic representation. In Proceedings of the. 24th Annual Conference of the Cognitive Science Society. (2002).

[8] Taylor Berg-Kirkpatrick, Dan Gillick, Dan Klein. Jointly Learning to Extract and Compress. ACL 2011, pp. 481–490 (2011)

[9] Aria Haghighi, Lucy Vanderwende. Exploring Content Models for Multi-Document Summarization. HLT-NAACL 2009, pp. 362–370 (2009)

[10] 北島理沙, 小林一郎. 文書内の事象を対象にした潜在的ディリクレ配分法による要約. DEIM 2011 (2011)

[11] Taku Kudo and Yuji Matsumoto, Japanese Dependency Analyisis using Cascaded Chunking, Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002), pp.63–69 (2002)

[12] Peng Li, Yinglin Wang, Wei Gao, Jing Jiang. Generating Aspect-oriented Multi-Document Summarization with Event-aspect model. EMNLP 2011, pp. 1137–1146 (2011)
[13] Hui Lin and Jeff Bilmes, A Class of Submodular Functions for Document Summarization, ACL 2011, pp. 510–520 (2011)
[14] 佐藤一誠, 中川裕志. Dirichlet Process Unigram Mixture Model に対する Collapsed 変分ベイズ法の適用, 情報処理学会論文誌, Vol.48 TOM19. pp.107-116, (2007)
[15] Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei. Hierarchical Dirichlet Processes. JASA 101(476): pp. 1566–1581, (2006)
[16] K. Yu, S. Yu, and V. Tresp, Dirichlet Enhanced Latent Semantic Analysis, AISTATS-05, (2005)