(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3921540号
(P3921540)
(45)発行日　平成19年5月30日(2007.5.30)　(24)登録日　平成19年3月2日(2007.3.2)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 210D
G06F 17/30 170A

請求項の数 12　(全 22 頁)

(21)出願番号　特願2003-290929(P2003-290929)
(22)出願日　平成15年8月8日(2003.8.8)
(65)公開番号　特開2005-63071(P2005-63071A)
(43)公開日　平成17年3月10日(2005.3.10)
審査請求日　平成15年8月8日(2003.8.8)

特許法第３０条第１項適用　平成１５年２月１３日鈴鹿サーキット　フラワーガーデンホテルにおいて開催された「情報アクセスのためのテキスト処理」シンポジウムで発表

(73)特許権者　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100130498
弁理士　佐野　禎哉
(72)発明者　野畑　周
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内
(72)発明者　井佐原　均
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内

最終頁に続く

(54)【発明の名称】文書セット分類装置及びそのプログラム

(57)【特許請求の範囲】
【請求項１】
任意に与えられた固有表現クラスの定義に基づき複数の文書の集合である文書セットに対して前記固有表現クラスの定義に基づいて得られる分類を付与するものであって、
前記文書セットに含まれる文書の中に出現する固有表現を抽出するとともに、抽出した固有表現が属する固有表現クラスを、固有表現と固有表現が属する固有表現クラスとが列挙されたデータ群を参照して判定する固有表現抽出手段と、
前記文書セットの主題が単独の固有表現に関するものか複数の固有表現に関するものかを判断し、かつ、該固有表現が任意に与えられた前記固有表現クラスのうち何れの固有表現クラスに属するかを判断する判断手段と、
前記判断手段が下した判断に基づき、前記文書セットの主題に係る固有表現が単独であるか複数であるか、及び、該固有表現が属している前記固有表現クラスという２つの要素より規定される分類についての情報を出力する出力手段と
を具備し、
前記判断手段は、
判断対象である固有表現の頻度が予め定められた閾値以上となる場合はその単独の固有表現に関するものと判断し、当該固有表現がなく且つ判断対象である固有表現クラスの頻度がそれぞれに対して予め定められた閾値以上となる場合は複数の固有表現に関するものと判断するものであり、かつ、
前記固有表現が、前記固有表現抽出手段の判定結果に従って、その判定された固有表現ク

ラスに属すると判断するものである
ことを特徴とする文書セット分類装置。
【請求項２】
前記文書セットに含まれる文書の中に出現するクラスタームを抽出するとともに、抽出したクラスタームが関連する固有表現クラスを、クラスタームとクラスタームが関連する固有表現クラスとが列挙されたデータ群を参照して判定するクラスターム抽出手段をさらに具備し、
前記判断手段は、
前記固有表現が、前記クラスターム抽出手段の判定結果に従って、その判定された固有表現クラスに属すると判断するものである
請求項１記載の文書セット分類装置。
【請求項３】
任意に与えられた固有表現クラスの定義に基づき複数の文書の集合である文書セットに対して前記固有表現クラスの定義に基づいて得られる分類を付与するものであって、
前記文書セットに含まれる文書の中に出現するクラスタームを抽出するとともに、抽出したクラスタームが関連する固有表現クラスを、クラスタームとクラスタームが関連する固有表現クラスとが列挙されたデータ群を参照して判定するクラスターム抽出手段と、
前記文書セットの主題が単独の固有表現に関するものか複数の固有表現に関するものかを判断し、かつ、該固有表現が任意に与えられた前記固有表現クラスのうち何れの固有表現クラスに属するかを判断する判断手段と、
前記判断手段が下した判断に基づき、前記文書セットの主題に係る固有表現が単独であるか複数であるか、及び、該固有表現が属している前記固有表現クラスという２つの要素より規定される分類についての情報を出力する出力手段と
を具備し、
前記判断手段は、
判断対象である固有表現の頻度が予め定められた閾値以上となる場合はその単独の固有表現に関するものと判断し、当該固有表現がなく且つ判断対象である固有表現クラスの頻度がそれぞれに対して予め定められた閾値以上となる場合は複数の固有表現に関するものと判断するものであり、かつ、
前記固有表現が、前記クラスターム抽出手段の判定結果に従って、その判定された固有表現クラスに属すると判断するものである
ことを特徴とする文書セット分類装置。
【請求項４】
前記判断手段は、前記文書セットに含まれる複数の文書の各々が作成若しくは発表された時点に関する情報を参照し、これら複数の文書のうちの一定の割合以上のものが予め定められた期間内に作成若しくは発表されていることを条件として、前記記事セットの主題に係る固有表現が単独でありかつその属する固有表現クラスがイベント名クラスである旨の判断を下す請求項１、２または３記載の文書セット分類装置。
【請求項５】
与えられた文書の中に存在するキーワードを抽出し、一の文書のキーワードと他の文書のキーワードとの類似度を算出し、その類似度が閾値を超える場合にこれらの文書を同一の文書セットに割り当てることを通じて、複数の文書から少なくとも一の文書セットを生成し得る文書セット生成手段をさらに具備する請求項１、２、３または４記載の文書セット分類装置。
【請求項６】
請求項１、２、３、４または５記載の文書セット分類装置とともに用いられるものであって、
前記文書セット分類装置が出力する、前記文書セットに付与された分類についての情報を参照し、この分類に対応した要約アルゴリズムにより前記文書セットに含まれる複数の文書を単一の文書に要約する要約手段を具備する文書要約装置。

【請求項７】
請求項１、２、４または５記載の文書セット分類装置を構成するために用いられるものであって、コンピュータを、少なくとも、
複数の文書の集合である文書セットに含まれる文書の中に出現する固有表現を抽出するとともに、抽出した固有表現が属する固有表現クラスを固有表現と固有表現が属する固有表現クラスとが列挙されたデータ群を参照して判定する固有表現抽出手段、
文書セットの主題が単独の固有表現に関するものか複数の固有表現に関するものかを判断し、かつ、該固有表現が任意に与えられた前記固有表現クラスのうち何れの固有表現クラスに属するかを判断する判断手段、及び、
前記判断手段が下した判断に基づき、前記文書セットの主題に係る固有表現が単独であるか複数であるか、及び、該固有表現が属している前記固有表現クラスという２つの要素より規定される分類についての情報を出力する出力手段
として機能させ、
前記判断手段は、前記コンピュータを、
判断対象である固有表現の頻度が予め定められた閾値以上となる場合はその単独の固有表現に関するものと判断し、当該固有表現がなく且つ判断対象である固有表現クラスの頻度がそれぞれに対して予め定められた閾値以上となる場合は複数の固有表現に関するものと判断し、かつ、
前記固有表現が、前記固有表現抽出手段の判定結果に従って、その判定された固有表現クラスに属すると判断するように機能させる
ことを特徴とするプログラム。
【請求項８】
前記コンピュータを、さらに
前記文書セットに含まれる文書の中に出現するクラスタームを抽出するとともに、抽出したクラスタームが関連する固有表現クラスを、クラスタームとクラスタームが関連する固有表現クラスとが列挙されたデータ群を参照して判定するクラスターム抽出手段として機能させ、
前記判断手段は、前記コンピュータを、
前記固有表現が、前記クラスターム抽出手段の判定結果に従って、その判定された固有表現クラスに属するとの判断を実行するものである請求項７記載のプログラム。
【請求項９】
請求項３、４または５記載の文書セット分類装置を構成するために用いられるものであって、コンピュータを、少なくとも、
複数の文書の集合である文書セットに含まれる文書の中に出現するクラスタームを抽出するとともに、抽出したクラスタームが関連する固有表現クラスを、クラスタームとクラスタームが関連する固有表現クラスとが列挙されたデータ群を参照して判定するクラスターム抽出手段と、
前記文書セットの主題が単独の固有表現に関するものか複数の固有表現に関するものかを判断し、かつ、該固有表現が任意に与えられた前記固有表現クラスのうち何れの固有表現クラスに属するかを判断する判断手段と、
前記判断手段が下した判断に基づき、前記文書セットの主題に係る固有表現が単独であるか複数であるか、及び、該固有表現が属している前記固有表現クラスという２つの要素より規定される分類についての情報を出力する出力手段と
として機能させ、
前記判断手段は、前記コンピュータを、
判断対象である固有表現の頻度が予め定められた閾値以上となる場合はその単独の固有表現に関するものと判断し、当該固有表現がなく且つ判断対象である固有表現クラスの頻度がそれぞれに対して予め定められた閾値以上となる場合は複数の固有表現に関するものと判断し、かつ、
前記固有表現が、前記クラスターム抽出手段の判定結果に従って、その判定された固有表

現クラスに属すると判断するように機能させる
ことを特徴とする文書セット分類装置。
【請求項１０】
前記判断手段は、前記文書セットに含まれる複数の文書の各々が作成若しくは発表された時点に関する情報を参照し、これら複数の文書のうちの一定の割合以上のものが予め定められた期間内に作成若しくは発表されていることを条件として、前記記事セットの主題に係る固有表現が単独でありかつその属する固有表現クラスがイベント名クラスである旨の判断を下す請求項７、８または９記載のプログラム。
【請求項１１】
さらにコンピュータを、与えられた文書の中に存在するキーワードを抽出し、一の文書のキーワードと他の文書のキーワードとの類似度を算出し、その類似度が閾値を超える場合にこれらの文書を同一の文書セットに割り当てることを通じて、複数の文書から少なくとも一の文書セットを生成し得る文書セット生成手段としても機能させる請求項７、８、９または１０記載のプログラム。
【請求項１２】
請求項６記載の文書要約装置を構成するために用いられるものであって、コンピュータを、少なくとも、
文書セット分類装置が出力する、文書セットに付与された分類についての情報を参照し、この分類に対応した要約アルゴリズムにより前記文書セットに含まれる複数の文書を単一の文書に要約する要約手段として機能させるプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、複数の文書を包含する文書セットに分類を付与するための分類装置及びそのプログラムに関する。
【背景技術】
【０００２】
複数文書の自動要約は、要約の研究において近年関心の高まっている分野である。米国のDocument Understanding Conference（DUC）や日本のText Summarization Challenge（TSC）でも、要約システムの評価を行う課題として複数文書の要約が対象に加えられている。複数文書要約とは、単一の主題について収集された複数の文書を含む文書セットを単一の文書に要約することである。より具体的に述べると、ある事件の始まりから終わりまでの一連の報告や、特定個人の行動・発言の内容、各地で発生した地震の被害状況等の主題に沿って収集された複数の文書より、当該主題に関する要約を生成することである。
【０００３】
要約の精度を向上させるためには、文書セットがもつ主題を正しく把握し、それに応じて適切な要約手法、出力形式を選択する必要があると考えられる。複数文書要約の観点から文書セットを分類する先行研究として、コロンビア大学のMcKeown等によるものがある（非特許文献１を参照）。McKeown等は、複数の新聞記事を包含する記事セットに付与すべき分類として、
（Ａ）Single-Event（特定の地域・期間に限定された単独の出来事に関する記事セット）
（Ｂ）Person-centered（特定人物にまつわる出来事を記述した記事セット）
（Ｃ）Multi-Event（異なる地域・期間にまたがった複数の出来事に関する記事セット。大抵は行動主体も異なる）
（Ｄ）Other（上記の３分類に当てはまらない、互いに漠然と関連している記事セット）
の４種類を定義した。そして、記事セットを分類する際の手がかりとして、記事セット中の全記事間のタイムスパン、同日に掲載された記事の割合、大文字で始まる語の頻度、he、she等の人称代名詞の頻度、を用いている。
【非特許文献１】K. R. McKeown and R. Barzilay and D. Evans and V. Hatzivassilogou and M. Yen Kan and B. Schiffman and S. Teufel, [online], “Columbia Multi-Docu

ment Summarization: Approach and Evaluation”, Online Proceedings of DUC2001 ＜http://www-nlpir.nist.gov/projects/duc/pubs/2001papers/columbia#redo.pdf＞

【発明の開示】

【発明が解決しようとする課題】

【０００４】

McKeown等の分類は、要約の対象となる記事セットによく見られる性質を効率よく分類している。しかしながら、幾つかの問題も抱えている。即ち、

・Otherに分類される記事セットが多くなる。これらに対しては他に適切な分類があるのではないかと考えられる。

・分類を判定するアルゴリズムに用いられている手がかりのうち、大文字で始まる語の頻度及びhe、she等の人称代名詞の頻度は英語に特化したものである。より一般的に記事セットの分類を行うためには、用いるべき手がかりを考慮する必要がある。

・上記非特許文献１で分類対象とされた記事データは、複数記事要約の評価ワークショップで使用するために作成されたものである。そのために、一般的な記事セットと比較して整えられているか、あるいは偏りが生じている。

・McKeown等の要約システムでは、実際に複数記事要約を行うときにMulti-EventとOtherとを同一視して同じ要約手法を適用しており、Multi-EventとOtherとを区別した意義が失われている。

【０００５】

以上に鑑みてなされた本発明は、特定言語の特性に依存せずなおかつ分類の網羅性を高めたより好適な分類を定義した上で、これに基づいた分類を文書セットに付与できる分類装置を提供するものである。

【課題を解決するための手段】

【０００６】

上述した課題を解決すべく、本発明では、任意に与えられた固有表現クラスの定義に基づき複数の文書の集合である文書セットに対し前記固有表現クラスの定義に基づいて得られる分類を付与するものとして、図１に示すように、前記文書セットに含まれる文書の中に出現する固有表現を抽出するとともに、抽出した固有表現が属する固有表現クラスを、固有表現と固有表現が属する固有表現クラスとが列挙されたデータ群を参照して判定する固有表現抽出手段（図示省略）と、前記文書セットの主題が単独の固有表現（Named Entity）に関するものか複数の固有表現に関するものかを判断し、かつ、該固有表現が何れの固有表現クラスに属するかを判断する判断手段１０１と、前記判断手段１０１が下した判断に基づき、前記文書セットの主題に係る固有表現が単独であるか複数であるか、及び、該固有表現が属している前記固有表現クラスという２つの要素より規定される分類についての情報を出力する出力手段１０２とを具備する文書セット分類装置を構成した。

【０００７】

本発明では、文書セットがもつ主題に関連する固有表現が属する固有表現クラスに基づいて分類を定義し、本発明に係る分類装置がこの定義された分類を文書セットに付与するものとした。このようなものであれば、特定言語に依存する度合いが低減するとともに、分類の網羅性が高まる、言い換えるならば文書セットがOtherに分類される可能性が小さくなる。

【０００８】

ここで、固有表現とは、文書中に含まれる人名、組織名等の固有名詞や日付、金額等の数値表現その他の情報抽出の要素となる表現を言う。固有表現は、情報として重要でありかつその内容を示す表現がほぼ一意に定まるものである。固有表現が属する固有表現クラスの定義は種々考えられるが、一態様として、人名（Person）、組織名（Organization）、地名（Location）、施設名（Facility）、固有物名（Product（製品名、法律名等））、イベント名（Event）の６種を内包するクラス定義を採用することができる。この固有表現クラス定義を採用した場合、文書セットに付与すべき分類は以下の通りとなる。即ち、

（Ａ）Single-Person（単一人物に関する文書セット）
（Ｂ）Single-Organization（単一組織に関する文書セット）
（Ｃ）Single-Location（単一地域に関する文書セット）
（Ｄ）Single-Facility（単一施設に関する文書セット）
（Ｅ）Single-Product（単一固有物に関する文書セット）
（Ｆ）Single-Event（単一イベントに関する文書セット）
（Ｇ）Multi-Person（複数人物に関する文書セット）
（Ｈ）Multi-Organization（複数組織に関する文書セット）
（Ｉ）Multi-Location（複数地域に関する文書セット）
（Ｊ）Multi-Facility（複数施設に関する文書セット）
（Ｋ）Multi-Product（複数固有物に関する文書セット）
（Ｌ）Multi-Event（複数イベントに関する文書セット）
（Ｍ）Other（その他）

但し、固有表現クラスの定義ひいては文書セットの分類の定義はこれに限定されない。よって、例えば、動物名クラス（Single/Multi-Animal）等を追加することができ、人名、動物名等を一のクラス、いわば行動主体を表現するクラスである「主体名クラス」にまとめることもできる。

【０００９】

文書セット分類装置における前記判断手段１０１は、ここで前記判断手段は、判断対象である固有表現の頻度が予め定められた閾値以上となる場合はその単独の固有表現に関するものと判断し、当該固有表現がなく且つ判断対象である固有表現クラスの頻度がそれぞれに対して予め定められた閾値以上となる場合は複数の固有表現に関するものと判断を実行し、かつ、前記固有表現が、前記固有表現抽出手段の判定結果に従って、その判定された固有表現クラスに属するとの判断を実行する。但し、ここに言う頻度は、文書セットに包含される複数の文書中に出現する特定の固有表現等の出現頻度には限られず、特定の固有表現等が出現する文書の数（文書セットにおける文書の頻度）であることがある。

【００１０】

加えて、文書セット分類装置は、文書セットに含まれる文書の中に出現するクラスタームを抽出するとともに、抽出したクラスタームが関連する固有表現クラスを、クラスタームとクラスタームが関連する固有表現クラスとが列挙されたデータ群を参照して判定するクラスターム抽出手段をさらに具備する構成として、前記判断手段１０１が、前記固有表現が、前記クラスターム抽出手段の判定結果に従って、その判定された固有表現クラスに属するとの判断を実行するものであってもよい。或いは文書セット分類装置を、前述した固有表現抽出手段、判断手段、出力手段を備える構成に代えて、クラスターム抽出手段、判断手段、出力手段を備える構成とすることもできる。なお、クラスタームとは、特定の固有表現クラスに関連の強い名詞または複合名詞のことである。例えば、「首相」等の役職名は人名クラスのクラスタームであり、「地震」等の名詞はイベント名クラスのクラスタームである。因みに、クラスタームは、固有表現そのものとは異なり、一般名詞である。

【００１１】

また、特に、文書セットに含まれる各文書の作成時または発表時が判明しているような場合において、前記判断手段１０１が、前記文書セットに含まれる複数の文書の各々の作成若しくは発表された時点に関する情報を参照し、これら複数の文書のうちの一定の割合以上のものが予め定められた期間内に作成若しくは発表されていることを条件として、前記記事セットの主題に係る固有表現が単独でありかつその属する固有表現クラスがイベント名クラスである旨の判断を下すものとしてもよい。このとき、当該文書セットにはSingle-Eventの分類が付与される。

【００１２】

図２に示すように、上記の文章セット分類装置が、与えられた文書の中に存在するキーワードを抽出し、一の文書のキーワードと他の文書のキーワードとの類似度を算出し、そ

の類似度が閾値を超える場合にこれらの文書を同一の文書セットに割り当てることを通じて、複数の文書から少なくとも一の文書セットを生成し得る文書セット生成手段１０３をさらに具備するものであれば、与えられた複数の文書を一または複数の文書セットに仕分けしこれに分類を付与するまでの処理を一括に実行可能となる。このものは、与えられた文書を基に一または複数の要約を自動生成するシステムを構築するために有用となる。

【００１３】

さらに、図３に示すように、上記の文書セット分類装置が文書セットに対して付与した分類を参照し、この分類に対応した要約アルゴリズムを選択して前記文書セットに含まれる複数の文書を要約する要約手段２０１を具備する文書要約装置を構成することで、より適切に複数文書の要約を実行することが可能となる。

【発明の効果】

【００１４】

以上に詳述した本発明によれば、特定言語の特性に依存せず、かつ分類の網羅性を高めたより好適な分類を文書セットに付与し得る。

【発明を実施するための最良の形態】

【００１５】

以下、本発明の一実施形態を、図面を参照して説明する。はじめに、本発明における分類の定義及びその妥当性について述べる。ここでは、分析対象とする文書セットとして複数の新聞記事を包含する記事セットを実験的に生成し、これを分析して固有表現クラスを基にした分類を定義する。そして、この分類を、テストデータとなる別の日本語新聞記事セット、及び、DUC2001で使用された英語新聞記事セットに適用することにより、分類の定義の妥当性を検証する。

【００１６】

記事セットの偏りを避けるため、日本語新聞記事コーパスから無作為に一の記事を抽出し、その記事に類似する記事を情報検索システムを使用して収集して、記事セットを生成した。具体的な手順は以下の通りである。

（１）日本語新聞記事コーパスに含まれる記事から無作為に記事を一つ選択する

（２）選択した記事からキーワード列を抽出する；キーワード列は、既知の日本語形態素解析ソフトウェア（黒橋禎夫、長尾真　日本語形態素解析システム Juman version 3.61. 京都大学, 1999.）を用いた形態素解析結果より時相名詞、副詞的名詞を除いた名詞のうち頻度２以上のものとした

（３）抽出したキーワード列を用いて、記事間の類似度を求める；各記事について上記（２）と同様にキーワード列を抽出し、キーワード同士の類似度をDice's coefficient（Diceの係数）を用いて求めた

（４）類似した記事を取り出す；同一の記事以外でDiceの係数が所定値（例えば、０．５）以上となる記事を類似記事と見なして取り出した。

【００１７】

以上の記事セット生成を複数回（例えば、５０回）繰り返して得られた複数（５０）記事セットのうち、３以上の記事を含む記事セットを選出すると２６記事セットとなった。これらから、さらに記事セットの内容がほぼ同じと考えられるものを省くと、１９記事セットが残った。これらの１９記事セットの主題を、図４に示す。

【００１８】

図４の分析結果に示されるように、Singleとは、記事セット中のほとんどの記事が単一のイベントや人名、組織名等の固有表現について記述してあるものである。他方、Multiとは、記事セット中の記事が複数の相異なるイベントや人名、組織名等の固有表現について記述してあるものである。

【００１９】

本発明では、文書セットたる記事セットの主題を、固有表現クラスのうち一つを選択することを通じて分類した。固有表現クラスの定義としては、拡張されたクラス定義（S. Sekine, K. Sudo and C. Nobata, “Extended named entity hierarchy”, In Proceeding

s of the LREC-2002 conference, 2002.）を採用した。この定義は階層的であるが、最上位の階層を用いてほぼ全ての記事セットの主題を分類することができた。固有表現クラスを割り当てず、Otherに分類した記事セットは一つだけである（図４における記事セットＮｏ．１４）。この記事セットに固有表現クラスを割り当てるとするならば動物名（Animal）クラスとなるが、動物名クラスに分類される記事セットはそれほど多くはないと考えられるので、Otherとした。上述したように固有表現クラスを選択的に割り当てた結果、既に言及した１３種類の分類が定義された。因みに、図４に示した分析結果には地名（Location）クラスの分類に対応する記事セットが存在しないが、特定の国や地域に関する記事セットは存在し得るし重要でもあると考えられるので、記事セットの分類の定義に含めた。

【００２０】

続いて、定義した分類に基づいてテストデータの分類を行う。ここで使用するテストデータは、日本語新聞記事コーパスから先に述べた方法と同様にして作成される。そして、テストデータとして作成された２０の記事セットについて、二人の被験者が独立に分類を付与した。原則として、被験者は各記事セットに一つの分類を割り当てたが、幾つかの記事セットについては複数の分類が可能であると判断して二つの分類を割り当てた。二人の被験者によって割り当てられた各分類の数を、図５に示す。被験者間の一致率は、被験者が最初に選択した分類同士を比較した場合には５５％、二番目の分類までを含めた場合には８５％であった。Otherに分類された記事セットはなかった。被験者が二つの分類を割り当てた記事セットの数は、それぞれ６と５であった。

【００２１】

被験者による分類結果を基に、テストデータの分類の正解データを作成した。被験者間で共通の分類となった１７記事セットではその割り当てられた分類を正解とする一方、被験者間で分類が分かれた３記事セットについては被験者同士の討論により正解の分類を決定した。なお、この正解データを用いて、後述する文書セット分類装置による自動分類の実験結果の評価を行う。

【００２２】

この分類の定義が他言語においても妥当であることを示すために、DUC2001の複数記事要約タスクで用いられたトレーニングデータである英語新聞記事セットに対しても分類の付与を試みた。先と同様に、二人の被験者が独立に記事セットに対して最適と判断した分類を付与した。被験者は各記事セットに対して一つないし二つの分類を付与した。二人の被験者によって割り当てられた各分類の数を、図６に示す。被験者が二つの分類を割り当てた記事セットはそれぞれ４セットあった。Otherの分類は、被験者の一人が二番目の分類として付与した１セットのみであった。被験者間の一致率は、被験者が最初に選択した分類同士を比較した場合に８０％、二番目の分類までを含めた場合は９３．３％であった。この一致率は、日本語記事のテストデータに対するものよりも高い。これは、DUC2001のデータが日本語記事テストデータよりも整えられており、記事セットの主題が意図的に選択されているからであると考えられる。

【００２３】

以降、定義された分類を記事セットに付与するための文書セット分類装置について詳述する。本実施形態における文書セット分類装置は、コンピュータ１に所定のプログラムをインストールすることで構成されるものである。コンピュータ１は、例えば、図７に示すように、プロセッサ１ａ、メインメモリ１ｂ、ハードディスクドライブに代表される補助記憶デバイス１ｃ等のハードウェア資源が、コントローラ１ｄ（即ち、いわゆるシステムコントローラ、Ｉ／Ｏコントローラ等）により制御され連携して動作するものである。また、図示しないが、電気通信回線を介して外部とのデータ授受を行うための通信デバイス、ユーザによる操作入力を受け付けるキーボードやポインティングデバイス等の入力デバイス、情報を画像ないし映像として表示するディスプレイ及びこのディスプレイに映像信号を送出するため表示制御デバイス（いわゆるグラフィクスチップ等）等を実装することを妨げない。

【００２４】
通常、プロセッサ１ａによって実行されるべきプログラムが補助記憶デバイス１ｃに格納されており、プログラムの実行の際には補助記憶デバイス１ｃからメインメモリ１ｂに読み込まれ、プロセッサ１ａによって解読される。そして、該プログラムに従い上記のハードウェア資源を作動して、少なくとも、判断手段１０１、出力手段１０２としての機能を発揮するようにしている。
【００２５】
判断手段１０１は、複数の記事（文書）を包含してなる記事セット（文書セット）の主題が単独の固有表現に関するものか複数の固有表現に関するものかを判断し、かつ、該固有表現が何れの固有表現クラスに属するかを判断する。入力として与えられる記事セットの要素である記事のデータは、通常、メインメモリ１ｂまたは補助記憶デバイス１ｃの所要の記憶領域に予め格納されている。よって、プロセッサ１ａが、プログラムに従い、メインメモリ１ｂまたは補助記憶デバイス１ｃに格納されている記事のデータを読み込み、これを基に記事セットに付与すべき分類についての判定を行う。
【００２６】
出力手段１０２は、前記判断手段１０１が下した判断より、前記記事セットに付与すべき分類についての情報、即ち、当該記事セットの主題に係る固有表現が単独であるか複数であるか及び該固有表現が属している固有表現クラスという２つの要素より規定される分類についての情報を出力する。情報の出力の態様としては、ディスプレイの画面への表示、プリンタ（図示せず）を使用したプリントアウト、通信デバイス及び電気通信回線を介した外部のコンピュータへの送信、メインメモリ１ｂまたは補助記憶デバイス１ｃへの書き込み、その他を挙げることができる。出力手段１０２の具体的構成は、記事セットに付与すべき情報の出力態様に応じたものとなる。
【００２７】
文書セット分類装置による分類の自動付与では、記事セットに含まれる複数の記事中に出現する単語や固有表現クラスの出現頻度と、記事頻度（記事セットにおいて所要の単語、固有表現クラス等が出現した記事の数。情報検索等で用いられるｉｄｆ値とは異なる）とを手がかりとして利用する。よって、コンピュータ１が、図８に示す固有表現抽出手段としての機能をも発揮し得ることが望ましい。固有表現抽出手段は、記事セットに含まれる記事の中に出現する固有表現を抽出するとともに、抽出した固有表現が属する固有表現クラスを判定する。固有表現抽出手段は、例えば、記事中の文章を単語に切り分けて品詞の付与を行う形態素解析ソフトウェア１０４ａと、形態素解析ソフトウェア１０４ａによる解析結果を参照して記事中に出現する固有表現を列挙しこれを固有表現クラス分けする固有表現抽出ソフトウェア１０４ｂとを用いて構成できる。形態素解析、固有表現抽出の一例を、図８に示している。図示例では、パターンベースのシステムで固有表現抽出を行っている。即ち、プロセッサ１ａが、プログラムに従い、入力として与えられた記事を形態素解析し、得られた形態素解析済み記事を固有表現リスト１０４ｃ（固有表現及びその属する固有表現クラスが列挙されたデータ群。通常、メインメモリ１ｂまたは補助記憶デバイス１ｃの所要の記憶領域に格納されている）に照らし合わせることで、記事中の固有表現を全て抽出する。しかる後、複数の固有表現が入れ子関係となっているもの（例えば、組織名クラスに属する固有表現「吉本工業」の中に、さらに人名クラスに属する固有表現「吉本」が存在）が存在しているときにはより文字列の長い固有表現を優先的に認定（即ち、「吉本」ではなく「吉本工業」という固有表現と認定）して固有表現を一意に決定し、その結果を出力する。判断手段１０１は、この固有表現抽出手段による出力を参照して、記事セットの分類を行うことができる。
【００２８】
また、固有表現に加え、記事中に出現するクラスタームも分類の手がかりとして用いることができる。クラスタームとは、特定の固有表現クラスに関連の強い名詞または複合名詞のことである。例示すると、「首相」等の役職名は人名クラスのクラスタームであり、「地震」等の名詞はイベント名クラスのクラスタームである。よって、コンピュータ１が

、図９に示すクラスターム抽出手段としての機能をも発揮し得ることが望ましい。クラスターム抽出手段は、記事セットに含まれる記事の中に出現するクラスタームを抽出するとともに、抽出したクラスタームが関連する固有表現クラスを判定する。クラスターム抽出手段は、例えば、記事中の文章を単語に切り分けて品詞の付与を行う形態素解析ソフトウェア１０４ａと、形態素解析ソフトウェア１０４ａによる解析結果を参照して記事中に出現するクラスタームを列挙しこれを固有表現クラス分けするクラスターム抽出ソフトウェア１０５ａとを用いて構成できる。形態素解析、クラスターム抽出の一例を、図９に示している。図示例もまた、上記の固有表現抽出の例と同様に、パターンベースのシステムでクラスターム抽出を行っている。即ち、プロセッサ１ａが、プログラムに従い、入力として与えられた記事を形態素解析し、得られた形態素解析済み記事をクラスタームリスト１０５ｂ（クラスターム及びその関連する固有表現クラスが列挙されたデータ群。通常、メインメモリ１ｂまたは補助記憶デバイス１ｃの所要の記憶領域に格納されている）に照らし合わせることで記事中のクラスタームを全て抽出し、その結果を出力する。判断手段１０１は、このクラスターム抽出手段による出力を参照して、記事セットの分類を行うことができる。因みに、クラスタームのリスト１０５ｂは、既存のシソーラスと人手による収集結果とから作成することができる。発明者が実験的に作成し使用しているクラスタームの数は約１６０００語である。

【００２９】

固有表現、クラスタームの他、記事の掲載日付や記事間のタイムスパン等もまた、記事セットの分類のための手がかりとして用いることが可能である。例えば、ある記事セットに含まれるほとんどの記事が同日かまたは所定の短い期間以内に掲載されたものであるならば、その記事セットをSingle-Eventに分類できる可能性が高い。

【００３０】

本実施形態における判断手段１０１が実行する判断のアルゴリズムに関して、図１０ないし図１４のフローチャートを参照して詳述する。本実施形態において、判断手段１０１は、下記の４つのアルゴリズムにより記事セットの分類を行う。まず、判断手段１０１は、第一のアルゴリズムに従い、入力として与えられた記事セットがSingle-Eventに分類されるか否かを判断する。この判断は、記事セットに含まれる複数の記事の各々が作成若しくは発表された時点に関する情報を参照して下される。記事セットに含まれる記事のうちの一定の割合以上のもの（「一定の割合」の設定値により、記事セットに含まれる記事のうち一部または全部の何れかが該当する）が、所定期間内に作成、公開、発表、掲載等されたものであるならば、判断手段１０１は当該記事セットをSingle-Eventに分類する。より具体的には、記事セットに含まれる記事が新聞記事である場合に、その大半が同日あるいは所定の短い期間内に掲載されたものであるならば当該記事セットにSingle-Eventの分類を付与する旨の判断を下す。このときの処理の手順を、図１０に示している。判断手段１０１は、入力として与えられた記事セットに含まれている各記事の掲載日付に関する情報（この情報は、例えば、記事データに関連づけてメインメモリ１ｂまたは補助記憶デバイス１ｃの所要の記憶領域に格納されている）を参照して（ステップＳ１０１）最も記事頻度の高い掲載日を確認し、この日に掲載された記事の数を計数する（ステップＳ１０２）。そして、この日に掲載された記事数またはこの日に掲載された記事数の記事セットに含まれる全記事数に対する割合が、予め設定された閾値T_aを上回っているならば（ステップＳ１０３）、与えられた記事セットをSingle-Eventに分類する（ステップＳ１０６）。また、記事セットに含まれる全記事の掲載日のタイムスパンを確認し（ステップＳ１０４）、このタイムスパンが予め設定された閾値T_sを下回っているならば（ステップＳ１０５）、与えられた記事セットをSingle-Eventに分類する（ステップＳ１０５）。つまり、特定の日に掲載された記事の割合、及び、記事間のタイムスパンの最大値の二つのパラメータを材料として判断を行う。記事セットをSingle-Eventに分類できなかったときには、第二のアルゴリズムに移行する。なお、記事中に出現する日付表現を参照することで第一のアルゴリズムを実行することを妨げない。

【００３１】

次に、判断手段１０１は、第二のアルゴリズムに従い、与えられた記事セットがSingle-class（何れかの固有表現クラス）に分類されるかどうか判断する。この判断は、記事頻度の高い固有表現について、その出現頻度、記事頻度を計数することで下される。特定の固有表現が記事セットに含まれる多くの記事にわたって頻繁に出現するならば、判断手段１０１はその固有表現を記事セットの主題を表すものと見なし、その固有表現が属する固有表現クラスを記事セットの分類の要素classとする。このときの処理の手順を、図１１に示している。判断手段１０１は、与えられた記事セットに対して前記固有表現抽出手段が行った固有表現抽出処理の結果出力を参照し（ステップＳ２０１）、記事セットに含まれる記事中に出現する固有表現及び固有表現が属する固有表現クラスのそれぞれについて、出現頻度を計数する（ステップＳ２０２）。かつ、記事中に出現する各固有表現についてその記事頻度を計数して、当該記事セットにおいて最も記事頻度の高い固有表現を選出する（ステップＳ２０３）。ここで、記事頻度が等しい複数の固有表現が存在する場合には、例えば出現頻度のより高い固有表現を選択する。しかして、選出された固有表現の記事頻度（または、選出された固有表現の記事頻度の記事セットに含まれる全記事数に対する割合）が予め設定された閾値T_dを上回っているならば（ステップＳ２０４）、この固有表現が属する固有表現クラスを記事セットの分類の要素classとすることができる。なお、記事セットをSingle-classに分類するに際し、さらなる判断処理を付加することを妨げない。即ち、図１１に示しているように、選出された固有表現の出現頻度／選出された固有表現が属する固有表現クラスの出現頻度、の比が予め定められた閾値T_wを上回っていることを条件として（ステップＳ２０７）、選出された固有表現が属する固有表現クラスを記事セットの分類の要素classとする（ステップＳ２０８）ものとしてもよい。但し、本実施形態では、トレーニングデータを調査した実験結果から、イベント名クラスに関しては他の固有表現クラスに優先して判断するものとした。従って、選出された固有表現がイベント名クラスに属するときには（ステップＳ２０５）、与えられた記事セットをそのままSingle-Eventに分類する（ステップＳ２０６）ようにしている。記事セットをSingle-classに分類できなかったときには、第三のアルゴリズムに移行する。

【００３２】

続いて、判断手段１０１は、第三のアルゴリズムに従い、与えられた記事セットがMulti-class（何れかの固有表現クラス）に分類されるかどうか判断する。この判断は、記事頻度の高い固有表現クラスについて、その出現頻度、記事頻度を計数することで下される。特定の固有表現クラスに属する固有表現が記事セットに含まれる多くの記事にわたって頻繁に出現するならば、判断手段１０１はその固有表現クラスを記事セットの主題を表すものと見なし、記事セットの分類の要素classとする。このときの処理の手順を、図１２に示している。判断手段１０１は、与えられた記事セットに対して前記固有表現抽出手段が行った固有表現抽出処理の結果出力を参照し（ステップＳ２０１）、記事セットに含まれる記事中に出現する固有表現及び固有表現が属する固有表現クラスのそれぞれについて、出現頻度を計数する（ステップＳ２０２。これらの処理は、既に第二のアルゴリズムにおいて実行されている）。かつ、記事中に出現する固有表現が属する各固有表現クラスについてその記事頻度を計数して、当該記事セットにおいて最も記事頻度の高い固有表現クラスを選出する（ステップＳ３０１）。ここで、記事頻度が等しい複数の固有表現クラスが存在する場合には、例えば出現頻度のより高い固有表現クラスを選択する。しかして、選出された固有表現クラスの記事頻度（または、選出された固有表現クラスの記事頻度の記事セットに含まれる全記事数に対する割合）が予め設定された閾値T_dを上回っているならば（ステップＳ３０２）、この固有表現クラスを記事セットの分類の要素classとすることができる。なお、記事セットをMulti-classに分類するに際し、さらなる判断処理を付加することを妨げない。即ち、図１２に示しているように、選出された固有表現クラスの出現頻度／全固有表現クラス（全固有表現）の出現頻度、の比が予め定められた閾値T_Cを上回っていることを条件として（ステップＳ３０５）、選出された固有表現クラスを記事セットの分類の要素classとする（ステップＳ３０６）ものとしてもよい。但し、本実施形態では、イベント名クラスに関しては他の固有表現クラスに優先して判断するも

のとし、選出された固有表現クラスがイベント名クラスに属するときには（ステップＳ３０３）、与えられた記事セットをそのままMulti-Eventに分類する（ステップＳ３０４）ようにしている。加えて、固有表現毎に相異する閾値T_Cを設定することを妨げない。例えば、選出された固有表現クラスの出現頻度／全固有表現クラスの出現頻度の値と比較される閾値T_Cについて、選出された固有表現クラスが地名クラス、組織名クラス、人名クラスの何れかである場合にはより厳しい即ちより大きい閾値T_{C1}を適用し、選出された固有表現クラスが上記以外である場合にはより緩い即ちより小さい閾値T_{C2}（$T_{C1} > T_{C2}$）を適用することができる。記事セットをMulti-classに分類できなかったときには、第四のアルゴリズムに移行する。

【００３３】

第三のアルゴリズムまでの過程で記事セットに付与すべき適切な分類を見出せなかった場合、判断手段１０１は、第四のアルゴリズムに従い、付与すべき分類を検討する。第四のアルゴリズムは、第二のアルゴリズムないし第三のアルゴリズムを、固有表現でなくクラスタームを対象として実行するものと言える。即ち、記事中に出現するクラスタームの頻度またはクラスタームが関連する固有表現クラスの頻度のうち少なくとも一方を材料として、与えられた記事セットの主題に係る固有表現が属する固有表現クラスの判断を下す。なお、特定のクラスタームが記事セット中の多くの記事にわたって頻繁に出現していても、当該記事セットに割り当てるべき分類はSingle-classでなくMulti-classとすることが望ましい。これは、クラスタームは固有名詞ではなく一般名詞であって、複数種の固有表現を指示し得るものであることによる。このときの処理の手順を、図１３及び図１４に示している。判断手段１０１は、与えられた記事セットに対して前記クラスターム抽出手段が行ったクラスターム抽出処理の結果出力を参照し（ステップＳ４０１）、記事セットに含まれる記事中に出現するクラスターム及びクラスタームが関連する固有表現クラスのそれぞれについて、出現頻度を計数する（ステップＳ４０２）。かつ、記事中に出現する各クラスタームについてその記事頻度を計数して、当該記事セットにおいて最も記事頻度の高いクラスタームを選出する（ステップＳ４０３）。ここで、記事頻度が等しい複数のクラスタームが存在する場合には、例えば出現頻度のより高いクラスタームを選択する。しかして、選出されたクラスタームの記事頻度（または、選出されたクラスタームの記事頻度の記事セットに含まれる全記事数に対する割合）が予め設定された閾値T_dを上回っているならば（ステップＳ４０４）、このクラスタームが関連する固有表現クラスを記事セットの分類の要素classとすることができる。なお、記事セットをMulti-classに分類するに際し、さらなる判断処理を付加することを妨げない。即ち、図１３に示しているように、選出されたクラスタームの出現頻度／選出されたクラスタームが関連する固有表現クラスの出現頻度、の比が予め定められた閾値T_wを上回っていることを条件として（ステップＳ４０７）、選出された固有表現が属する固有表現クラスを記事セットの分類の要素classとする（ステップＳ４０８）ものとしてもよい。但し、イベント名クラスに関しては他の固有表現クラスに優先して判断するものとし、選出されたクラスタームがイベント名クラスに関連するものであるときには（ステップＳ４０５）、与えられた記事セットをそのままMulti-Eventに分類する（ステップＳ４０６）ようにしている。上記に加えて、記事中に出現するクラスタームが関連している各固有表現クラスについてその記事頻度を計数し、当該記事セットにおいて最も記事頻度の高い固有表現クラスを選出する（ステップＳ４０９）。記事頻度が等しい複数の固有表現クラスが存在する場合には、例えば出現頻度のより高い固有表現クラスを選択する。しかして、選出された固有表現クラスの記事頻度（または、選出された固有表現クラスの記事頻度の記事セットに含まれる全記事数に対する割合）が予め設定された閾値T_dを上回っているならば（ステップＳ４１０）、この固有表現クラスを記事セットの分類の要素classとすることができる。記事セットをMulti-classに分類するに際しては、さらなる判断処理を付加することができる。即ち、図１４に示しているように、選出された固有表現クラスの出現頻度／全固有表現クラス（全クラスターム）の出現頻度、の比が予め定められた閾値T_Cを上回っていることを条件として（ステップＳ４１３）、選出された固有表現クラスを記事セットの分類の要素classと

する（ステップＳ４１４）ものとできる。但し、イベント名クラスに関しては他の固有表現クラスに優先して判断するものとし、選出された固有表現クラスがイベント名クラスに属するときには（ステップＳ４１１）、与えられた記事セットをそのままMulti-Eventに分類する（ステップＳ４１２）ようにしている。なお、ここでも、第三のアルゴリズムと同様、固有表現毎に相異する閾値T_{C1}、T_{C2}を設定することが許容される。

【００３４】

上記の全てのアルゴリズムを用いても分類を付与できなかった場合、判断手段１０１は、予め定められたデフォルトの分類を当該記事セットに付与する（ステップＳ４１５）。デフォルトの分類は、例えば、Multi-EventまたはOtherとする。

【００３５】

上述のテストデータに対し、本実施形態の文書セット分類装置を使用して分類を付与する自動分類実験を行った結果について述べる。なお、アルゴリズム中の各閾値の決定は、ここではトレーニングデータを基に人手で行う。各閾値の設定は、T_a＝０．３３、T_s＝１５０、T_d＝０．９０、T_w＝０．４０、T_{C1}＝０．８０、T_{C2}＝０．４０とした。但し、閾値の大きさがここに示す値に限られないことは言うまでもない。テストデータに対する自動分類実験の結果の評価を、図１５に示す。図１５には、被験者による分類付与の結果の評価及びベースラインをも示した。被験者の評価は、各被験者が付与した分類と正解との比較評価である。両被験者の正解に対する評価は両被験者間の一致率５５％よりも高いが、これは分類の正解が両被験者による分類付与結果を総合して作成されたためである。ベースラインは、トレーニングデータにおいて最も頻度の高い分類（この実験では、Single-Event）の記事セットのテストデータにおける数（及び、占める割合）である。文書セット分類装置について、「一致」の値は文書セット分類装置が出力した分類が正解に一致した記事セット数（及び、割合）を示し、「部分一致」の値は文書セット分類装置が出力した分類が被験者によって付与された分類の何れかに一致した数（及び、割合）を示す。被験者が複数の分類を付与した記事セットに関してはその双方を含む。被験者について、「一致」の値は被験者が最初に与えた分類が正解の分類に一致した記事セット数（及び、割合）を示し、「部分一致」の値は被験者が二番目に与えた分類も含めて正解の分類に一致した記事セット数（及び、割合）を示す。

【００３６】

文書セット分類装置は、２０記事セットのうち９つを正しく分類し、さらに２つの記事セットについてはその分類結果が被験者が与えた分類に含まれていた。分類が正しくなかった残り９記事セットのうちの３つは、正解の分類がSingle-Productであるのに対してSingle-Eventと分類していた。実験に使用されたこれら記事セットの中に現れる固有物名（Product）は、特定の法案や国際条約等であり、記事セットに含まれる記事はその法案の審議や国際条約に対する発言について記述されたものであった。現在のアルゴリズムでは、Single-Eventを優先して分類するようになっているため、このような誤りが生起したと考えられる。しかしながら、正解の分類に関連する法案や国際条約等の固有物名は記事セット中の記事全体にわたって現れているため、判断手段１０１が一旦Single-Eventの分類を与えておきながらその後の判断過程を継続し、Single-Productの分類を与え直すことができるように構成することは可能であると考えられる。

【００３７】

別の３記事セットでは、正解の分類がSingle-Eventであるのに対して異なる分類を付与していた。これらの記事セットでは、イベント名（Event）にあたる記述が固有表現ではなく、句や節の形で表されていた。一例を挙げると、「クリントン前大統領のホワイトハウス元実習生モニカ・ルインスキさんに対する不倫疑惑」という表現は固有表現ではないが、特定のイベントを指す表現である。現状の固有表現抽出システムでは、このような表現を一の固有表現として認識することはできない。また、記事セットのタイムスパンは設定した閾値T_sよりも大きかった（上記例では、１年以上）ため、判断手段１０１は当該記事セットにSingle-Eventの分類を付与することができなかった。このことは、Single-Eventの分類を確実に付与するためには現在用いている手がかりの他に新たな手がかりを用

いる必要があると言うことを示唆している。

【００３８】

また、実験の過程で、記事セットの中には本質的に一以上の分類を付与し得るものがあることが分かった。その理由の一つは、Single-Event、Multi-Event等、イベント名に基づく分類を付与すべき記事セットには他の固有表現クラスに基づく分類を付与可能なケースも多いことである。イベントの多くは、特定の人名や組織名、地名等に関連することがしばしばであり、イベントに関する記事を集めた記事セットに対しそのイベントに関連する（イベント名以外の）固有表現に焦点を絞ることが可能である。もう一つの理由として、イベントにおけるSingleとMultiとの区別が難しいことが挙げられる。あるイベントの中には、幾つかの小さなイベントが包含されることがある。例えば、「シドニーオリンピック」に関する記事セットは、一つのスポーツイベントを対象とするものとしてSingle-Eventの分類を付与し得るが、この記事セット中に複数の種目の結果を報じる記事が含まれているならば、それらに着目することでMulti-Eventの分類を付与することも可能である。イベントの単位をどのように認識するかは、被験者の観点に依存する。

【００３９】

因みに、上述の実験では、固有表現抽出手段による結果出力に人手による修正を加えて固有表現抽出タスクにおける誤りを排除している。本実施形態の文書セット分類装置を用いて機械的に分類を付与するにあたり、固有表現抽出タスクの段階でエラーが生起することを完全に避けるのは難しい（完璧な固有表現抽出ソフトウェアは現存しないため）。固有表現抽出手段における固有表現抽出処理を補完するためには、共参照や、文献（Y. Shin-yama, S. Sekine K. Sudo, and R. Grishman, “Automatic paraphrase acquisition from news articles”, In Proceedings of the HLT-2002 conference, 2002.）に述べられているようなイベントの記述に関する言い換え表現の認識手法等を導入することが考えられる。

【００４０】

以上では、入力として与えられる複数の記事（文書）が予め記事セット（文書セット）に仕分けされていることを前提としていた。しかしながら、入力として複数の記事が単純に与えられるような状況も考えられる。このような場合において、文書セット分類装置が、与えられる複数の記事を一または複数の記事セットに仕分けし、仕分けした記事セットに分類を付与するまでの処理を機械的に実行し得ることが好ましい。即ち、文書セット分類装置を構成するコンピュータ１が、図２に示す文書セット生成手段１０３としての機能をも発揮し得ることが好ましい。

【００４１】

文書セット生成手段１０３は、ソフトウェアを主体として構成され、入力として与えられた文書の中に存在するキーワードを抽出し、一の文書のキーワードと他の文書のキーワードとの類似度を算出し、その類似度が閾値を超える場合にこれらの文書を同一の文書セットに割り当てることを通じて、複数の文書から少なくとも一の文書セットを生成する処理を行う。文書セット生成手段１０３が実行する処理の手順は、既に述べた記事セットの生成手法に類似する。即ち、プロセッサ１ａが、プログラムに基づき、入力として与えられた記事データ（通常、メインメモリ１ｂまたは補助記憶デバイス１ｃの所要の記憶領域に格納されている）のうちの一つを選択的に読み込み、この記事データよりキーワード列を抽出する。キーワードの抽出は、形態素解析ソフトウェアを利用して行うことができる。例えば、記事データを形態素解析した結果より時相名詞、副詞的名詞を除いた名詞のうち頻度が所定値（例えば、２）以上のものをキーワードとして抽出する。入力として与えられた各記事データについて上記の方法でキーワードを抽出した後、プロセッサ１ａが、一の記事データに係るキーワード列と他の記事データに係るキーワード列との間の類似度を算出する。プロセッサ１ａが算出する類似度の指標としては、Dice's coefficient、Jacquard measure、cosine similarity等を採用することができる。その上で、類似度の指標が所定値（例えば、０．５）以上である複数の記事を類似記事として一の記事セットに含めることを通じて、記事セットの生成を行う。

【００４２】
ところで、文書セット分類装置が出力する記事セットの分類についての情報を参照することで、当該記事セットに適した要約アルゴリズムを選択し得る。よって、文書セット分類装置が出力する分類についての情報を参照し、この分類に対応した要約アルゴリズムを用いて記事セット（文書セット）に含まれる複数記事（文書）の要約を行う要約装置を構築すれば、要約の精度の向上を図ることができる。
【００４３】
本実施形態における文書要約装置は、文書セット分類装置を構成するコンピュータ１またはこのコンピュータ１とは別のコンピュータ（図示せず）に所定のプログラムをインストールすることで構築される。通常、プログラムは補助記憶デバイス１ｃに格納され、その実行の際には補助記憶デバイス１ｃからメインメモリ１ｂに読み込まれてプロセッサ１ａにより解読される。そして、該プログラムに従いハードウェア資源を作動して、図３に示す要約手段２０１としての機能を発揮するようにしている。
【００４４】
要約手段２０１は、文書セット分類装置が出力する、記事セットに付与された分類についての情報を参照し、この分類に対応した要約アルゴリズムにより前記記事セットに含まれる複数の記事を単一の文書に要約する。即ち、プロセッサ１ａが、プログラムに基づき、記事セットに付与された分類についての情報を参照し、この分類に対応した要約アルゴリズムを選択する。しかる後、入力として与えられた記事セットに含まれる記事のデータを読み込み、要約の生成を行う。本実施形態における要約手段２０１は、ソフトウェアを主体として構成される。要約手段２０１の主体となるソフトウェアには、既知の複数記事要約ソフトウェアを応用できる。既知の複数記事要約ソフトウェアでは、一般に、記事セットに含まれる複数記事から重要と判断される文（ないし、文章）を抽出し、抽出した文を基にして要約を生成する。それぞれの文の重要度は、文の位置、文の長さ、文中に出現する単語の頻度、見出しとの類似度等の複数の条件に関するスコアを加算して算定される。また、重要度の算定にあたっては、個々の条件に関するスコアに対して重み付けがなされる（重みは、トレーニングデータを用いた訓練を通じて得られる）。しかして、本実施形態における要約手段２０１では、文書セット分類装置によって付与された分類に基づくスコアを加味して、それぞれの文の重要度を算出することとしている。具体例を挙げると、対象の記事セットに付与された分類がSingle-Personでありその主題を表す固有表現が「小泉」である場合には、各記事において、人名として認識できる「小泉」を含む文にスコアを与える。あるいは、対象の記事セットに付与された分類がMulti-Organizationである場合には、各記事において、組織名を含む文にスコアを与える。このように、分類に基づくスコアを加味してそれぞれの文の重要度を算定することにより、生成される要約の的確性の向上が期待できる。
【００４５】
加えて、要約手段２０１が、複数記事中の一の文と他の文との間の類似度を（共通する単語の個数等を参照することで）算出して類似する複数の文を抽出し、抽出した類似する文のうち要約生成に用いる文を選出するものとしてもよい。その上で、類似する複数の文より要約生成に用いる文を選出するための処理を、対象の記事セットに付与された分類に応じて変更することが好適である。具体例を挙げると、対象の記事セットに付与された分類がSingle-classである場合には、類似する複数の文が同一の事物を表現している可能性が高いことから、類似する複数の文のうちの一部の文のみを代表として選出する。つまり、類似する文として、「京都市で起きた震度４の地震で、３人が軽いけがを負った。」、「京都市で起きた震度４の地震で、新たに２人が入院し、けが人は５人となった。」というような複数の文が抽出されたとき、これらのうち何れか一文のみを要約記事の要素として選出する。これらの文のうちの何れを選択するかは、それぞれの文の重要度のスコアを参照する、時系列で最も後者の文を選択する等のヒューリスティクスにより決定できる。他方、対象の記事セットに付与された分類がMulti-classである場合には、類似する複数の文が相異なる事物を表現している可能性が高いことから、重要度スコアの高い文に類似

する一部または全部の文をまとめて選出する。つまり、類似する文として、「京都市で起きた震度４の地震で、３人が軽いけがを負った。」、「大阪市で起きた震度５の地震で、５人が入院し、８人が軽いけがを負った。」というような複数の文が抽出されたとき、これらの文は表現上似ているものの相異なるイベントを記述していると考えられる。であるから、これらの文の全てを要約記事の要素として選出することもあり得る。

【００４６】

総じて言えば、本実施形態における文書要約装置の要約手段２０１は、記事セットに付与された分類に基づく重要度スコアを加味してそれぞれの文の重要度を算定するプロセス、及び／または、記事セットに付与された分類に応じて類似する複数の文の取捨選択の手法を変えるプロセスを、既存の複数記事要約ソフトウェアに追加したものとして構成可能である。そして、対象の記事セットに付与された分類に応じて異なる要約アルゴリズムの要約処理を実行可能である。因みに、Single-Personの分類が付与された記事セットより伝記的な記述を要約出力させたり、Multi-Productの分類が付与された記事セットより製品の名称・機能・値段等の要素を抽出させて表の形態で出力させたりというように、記事セットの分類に応じた多様な要約を要約手段２０１に出力させることも考えられる。

【００４７】

本実施形態によれば、複数の文書の集合である文書セットに対し分類を付与するものとして、前記文書セットの主題が単独の固有表現（Named Entity）に関するものか複数の固有表現に関するものかを判断し、かつ、該固有表現が何れの固有表現クラスに属するかを判断する判断手段１０１と、前記判断手段１０１が下した判断に基づき、前記文書セットの主題に係る固有表現が単独であるか複数であるか、及び、該固有表現が属している固有表現クラスという２つの要素より規定される分類についての情報を出力する出力手段１０２とを具備する文書セット分類装置を構成したため、網羅性の高い分類を文書セットに付与可能となる。

【００４８】

前記判断手段１０１が、前記文書セットに含まれる複数の文書の中に出現する固有表現の頻度または固有表現クラスの頻度のうち少なくとも一方を材料として、前記文書セットの主題に係る固有表現が単独であるか複数であるかの判断及び該固有表現が属する固有表現クラスの判断を実行するため、大文字で始まる語の頻度やhe、she等の人称代名詞の頻度等の特定言語の特性に依存することなく分類を実行し得る。即ち、より一般的に記事セットの分類を行うことができる。

【００４９】

前記判断手段１０１が、前記文書セットに含まれる複数の文書の中に出現するクラスタームの頻度またはクラスタームが関連する固有表現クラスの頻度のうち少なくとも一方を材料として、前記文書セットの主題に係る固有表現が属する固有表現クラスの判断を実行するため、固有表現のみを手がかりとして分類できない文書セットに対しても適切な分類を付与することが可能である。

【００５０】

また、前記判断手段１０１が、前記文書セットに含まれる複数の文書の各々の作成若しくは発表された時点に関する情報を参照し、これら複数の文書のうちの一部または全部が予め定められた期間内に作成若しくは発表されていることを条件として、前記記事セットの主題に係る固有表現が単独でありかつその属する固有表現クラスがイベント名クラスである旨の判断を下すものとしており、少なくともSingle-Eventクラスの文書セットを速やかに分類できる。

【００５１】

文章セット分類装置が、与えられた文書の中に存在するキーワードを抽出し、一の文書のキーワードと他の文書のキーワードとの類似度を算出し、その類似度が閾値を超える場合にこれらの文書を同一の文書セットに割り当てることを通じて、複数の文書から少なくとも一の文書セットを生成し得る文書セット生成手段１０３をさらに具備するものであれば、与えられた複数の文書を一または複数の文書セットに仕分けしこれに分類を付与する

までの処理を一括に実行可能となる。このものは、与えられた文書を基に一または複数の要約を自動生成するシステムを構築するために有用となる。

【００５２】

さらに、上記の文書セット要約装置が文書セットに対して付与した分類を参照し、この分類に対応した要約アルゴリズムを選択して前記文書セットに含まれる複数の文書を要約する要約手段２０１を具備する文書要約装置を構成して、より適切な複数文書の要約処理を可能とすることができる。

【００５３】

なお、本発明は以上に詳述した実施形態に限られるものではない。特に、本発明で定義した記事セットの分類やその分類を行う文書セット分類装置及びそのプログラムは、自動要約以外での応用も考えられる。例示すると、情報検索において、検索された結果中の上位の記事を用いた検索後の再ランク付けや検索結果の効率的な表示を行うことができる。さらに、オープンドメインの情報抽出に利用することも考えられる。従来の情報抽出ではドメインが限定されており、記事の主題や分類は前提として与えられていた。しかし、ドメインを限定することなく情報抽出を行うためには、対象となるドメインの情報、即ち記事セットの分類を動的に実施する必要があると考えられるからである。

【００５４】

その他各部の具体的構成や図１０ないし図１４に示す処理の手順等もまた、上記実施形態に限られるものではなく、本発明の趣旨を逸脱しない範囲で種々変形が可能である。勿論、パーソナルコンピュータその他の汎用的な情報処理装置にプログラムをインストールして本発明に係る文書セット分類装置を構成することが可能であって、専用の装置を製造することが必須であるわけではない。

【図面の簡単な説明】

【００５５】

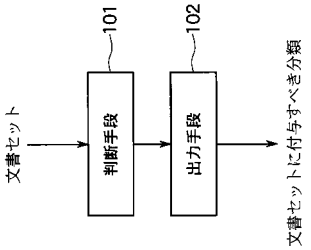
【図１】本発明に係る文書セット分類装置の構成説明図。
【図２】本発明に係る文書セット分類装置の構成説明図。
【図３】本発明に係る文書セット分類装置の構成説明図。
【図４】複数記事を包含する記事セットを例示する図。
【図５】２人の被験者による記事分類実験の結果を示す図。
【図６】２人の被験者による記事分類実験の結果を示す図。
【図７】本発明の一実施形態における文書セット分類装置が具備するハードウェア資源を示す図。
【図８】固有表現抽出手段による固有表現抽出処理を説明する図。
【図９】クラスターム抽出手段によるクラスターム抽出処理を説明する図。
【図１０】判断手段が実行する判断処理の手順を示すフローチャート。
【図１１】同フローチャート。
【図１２】同フローチャート。
【図１３】同フローチャート。
【図１４】同フローチャート。
【図１５】文書セット分類装置による自動分類実験の結果を示す図。

【符号の説明】

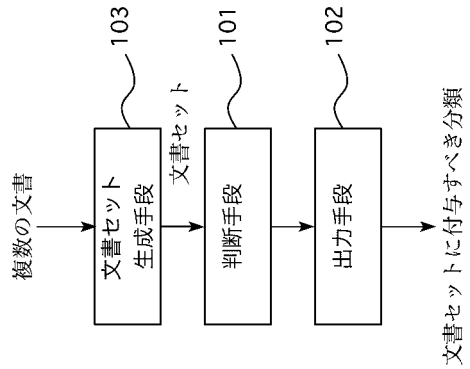
【００５６】

１…コンピュータ（文書セット分類装置、文書要約装置）
１０１…判断手段
１０２…出力手段
１０３…文書セット生成手段
２０１…要約手段

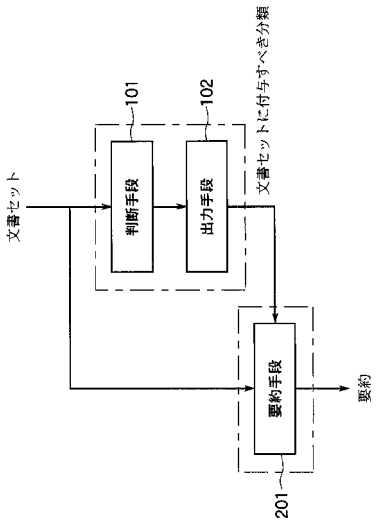
【図１】



【図２】



【図３】



【図４】

記事セットNo.	記事数	記事セットの主題	分類
1	14	ロシアのチェチェンとの紛争	Single-Event
2	10	日本各地の地震速報	Multi-Event
3	8	愛犬家連続失踪事件	Single-Event
4	7	民主リベラル新党の結成	Single-Event
5	6	書き初め大会、競技大会等	Multi-Event
6	6	各国の元首の動向	Multi-Person
7	6	日米首脳会談	Single-Event
8	4	大相撲の記事。特に、貴乃花について	Single-Person
9	4	伊東議員の北海道知事選出馬	Single-Person
10	4	インドネシア・ボルトガル首脳会談	Single-Event
11	3	浄土真宗本願寺派の記事	Single-Organization
12	3	市立船橋の高校サッカー大会での活躍	Single-Organization
13	3	ビールメーカー各首脳のコメント	Multi-Organization
14	3	動物が関与した事件の記事	Other
15	3	レコード大賞、ノーベル賞に関する記事	Multi-Artifact
16	3	中央アルプス・木曽駒ヶ岳付近で発生した雪崩	Single-Event
17	3	関西国際空港に関する記事	Single-Location
18	3	サッカーのインターコンチネンタル選手権	Single-Event
19	3	野茂投手の大リーグ挑戦	Single-Person

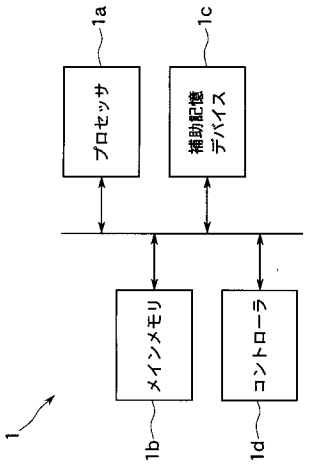
【図５】

分類	被験者A	被験者B
Single-Person	2	1
Single-Organization	2	
Single-Product	1	2
Single-Event	7	7
Multi-Location	1	
Mulit-Product	1	1
Multi-Event	6	9

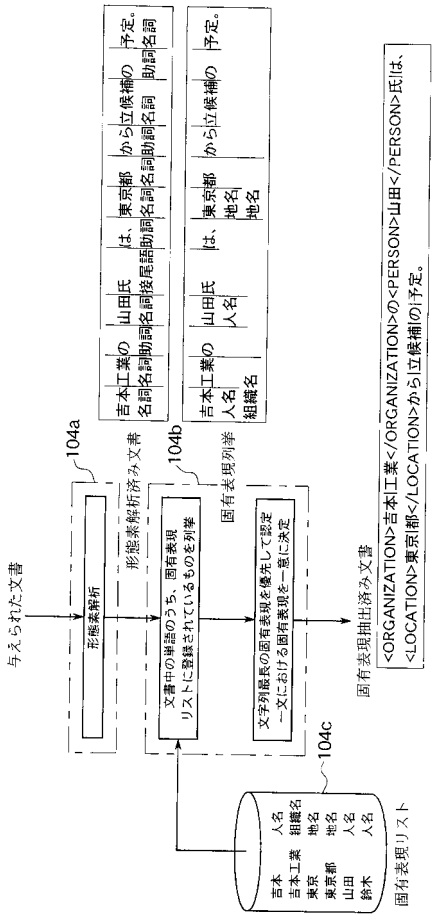
【図６】

分類	被験者A	被験者B
Single-Person	12	10
Single-Location	2	4
Single-Product	3	5
Single-Event	6	6
Multi-Event	7	5

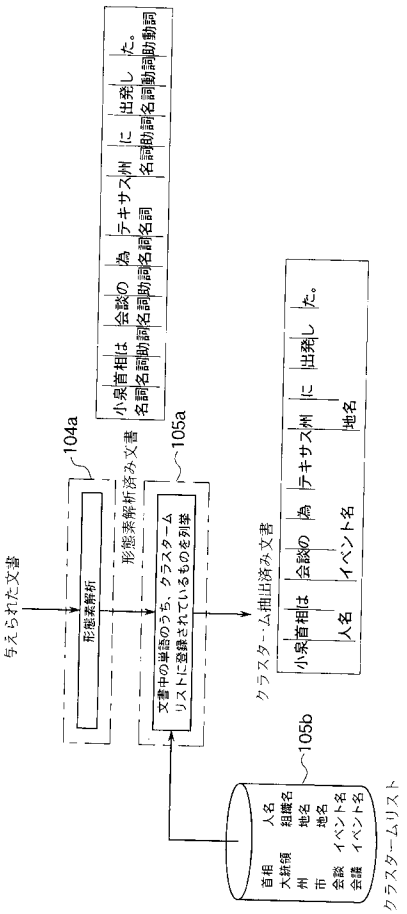
【図７】


1a
プロセッサ
1c
補助記憶デバイス
メインメモリ
コントローラ
1
1b
1d

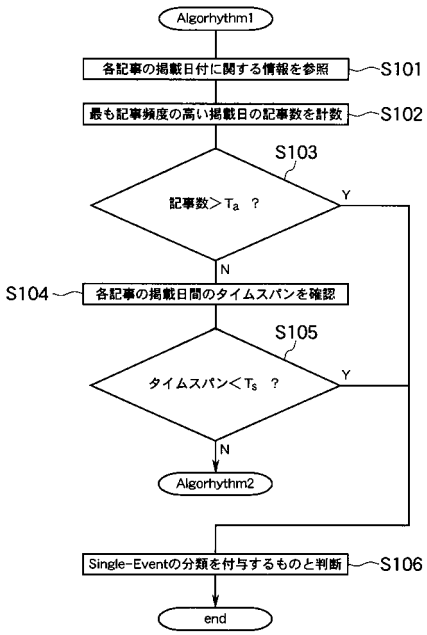
【図８】


与えられた文書
形態素解析
104a
形態素解析済み文書
文書中の単語のうち、固有表現リストに登録されているものを列挙
104b
固有表現列挙
文字列最長の固有表現を優先して認定
一文における固有表現を一意に決定
104c
固有表現抽出済み文書
<ORGANIZATION>吉本工業</ORGANIZATION>の<PERSON>山田</PERSON>氏は、<LOCATION>東京都</LOCATION>から立候補の予定。
固有表現リスト

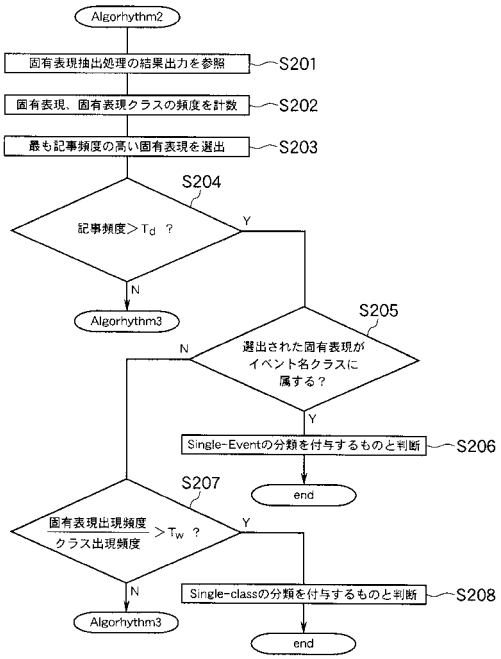
【図９】


与えられた文書
形態素解析
104a
形態素解析済み文書
文書中の単語のうち、クラスターム
リストに登録されているものを列挙
105a
クラスターム抽出済み文書
105b
クラスタームリスト

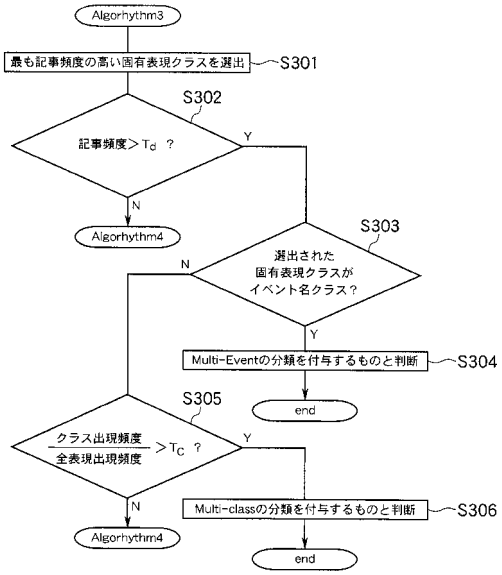
【図１０】


Algorhythm1
各記事の掲載日付に関する情報を参照
S101
最も記事頻度の高い掲載日の記事数を計数
S102
S103
記事数＞Ta ？
Y
N
S104
各記事の掲載日間のタイムスパンを確認
S105
タイムスパン＜Ts ？
Y
N
Algorhythm2
Single-Eventの分類を付与するものと判断
S106
end

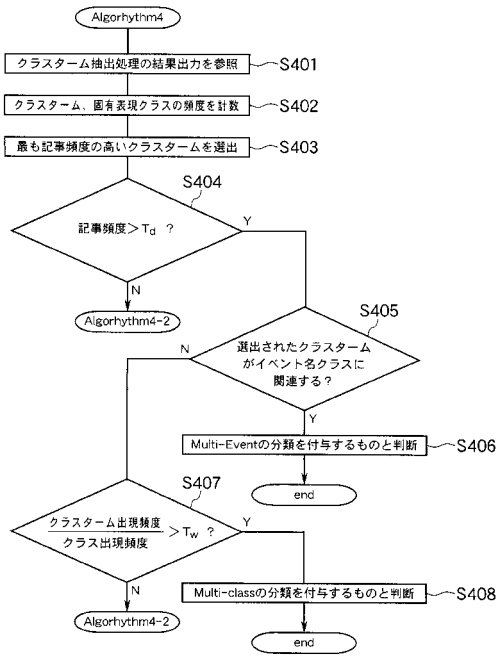
【図１１】



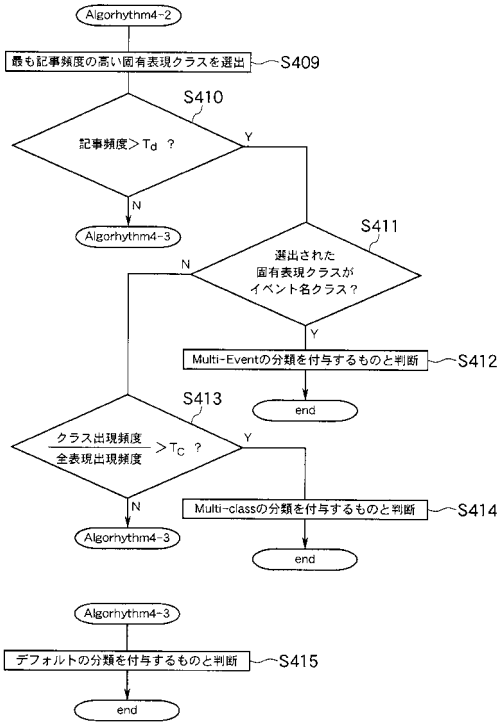
【図１２】



【図１３】



【図１４】



【図１５】

	一致	部分一致	誤り
文書セット分類装置	9(45%)	11(55%)	9
被験者A	14(70%)	17(85%)	3
被験者B	15(75%)	18(90%)	2
ベースライン	7(35%)	7(35%)	13

フロントページの続き

(72)発明者 関根 聡
アメリカ合衆国 １０００３ ニューヨーク州 ニューヨーク ブロードウェイ ７１５ セブンスフロア ニューヨーク大学内

審査官 丹治 彰

(56)参考文献 特開２００１－３３１５２９（ＪＰ，Ａ）
特開平０４－１０６６６３（ＪＰ，Ａ）
Kathleen R. McKeown et al., Columbia Multi-Document Summarization: Approach and Evaluation, Proceedings of DUC2001, ２００１年 9月１４日, [平成１８年5月１７日検索], ＵＲＬ, http://www-nlpir.nist.gov/projects/duc/pubs/2001papers/columbia_redo.pdf
難波英嗣ほか, テキスト自動要約－知的活動支援の基本技術として－ １ ここまで来たテキスト自動要約, 情報処理, 日本, 社団法人情報処理学会, ２００２年１２月１５日, 第４３巻 第１２号, 第１２８７頁～第１２９４頁

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ １７／３０
ＪＳＴＰｌｕｓ（ＪＤｒｅａｍ２）