

(19) 世界知的所有権機関
国際事務局(43) 国際公開日
2007年9月27日 (27.09.2007)

PCT

(10) 国際公開番号
WO 2007/108529 A1

- (51) 国際特許分類:
G06F 17/30 (2006.01) G06F 17/21 (2006.01)
- (21) 国際出願番号: PCT/JP2007/055958
- (22) 国際出願日: 2007年3月23日 (23.03.2007)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願2006-081598 2006年3月23日 (23.03.2006) JP
- (71) 出願人 (米国を除く全ての指定国について): 日本電気株式会社 (NEC CORPORATION) [JP/JP]; 〒1080014 東京都港区芝五丁目7番1号 Tokyo (JP).
- (72) 発明者; および
- (75) 発明者/出願人 (米国についてのみ): 水口 弘紀 (MIZUGUCHI, Hironori) [JP/JP]; 〒1080014 東京都港区芝五丁目7番1号 日本電気株式会社内 Tokyo (JP).

土田 正明 (TSUCHIDA, Masaaki) [JP/JP]; 〒1080014 東京都港区芝五丁目7番1号 日本電気株式会社内 Tokyo (JP). 久寿居 大 (KUSUI, Dai) [JP/JP]; 〒1080014 東京都港区芝五丁目7番1号 日本電気株式会社内 Tokyo (JP). 河合 英紀 (KAWAI, Hideki) [JP/JP]; 〒1080014 東京都港区芝五丁目7番1号 日本電気株式会社内 Tokyo (JP).

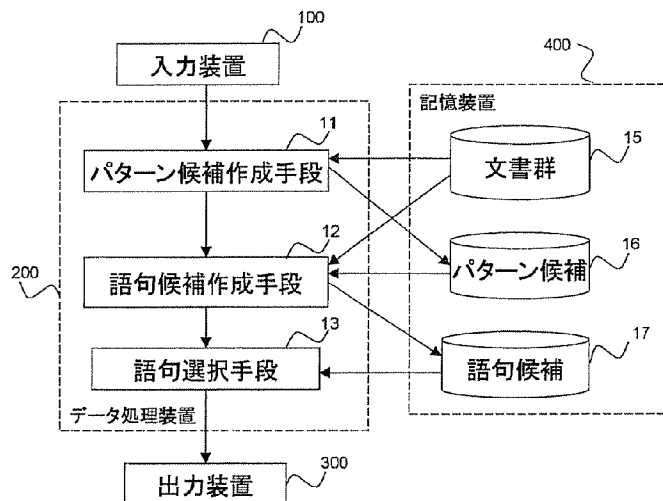
(74) 代理人: 山下 穰平 (YAMASHITA, Johei); 〒1050001 東京都港区虎ノ門五丁目13番1号虎ノ門4 O M T ビル 山下国際特許事務所 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, MG,

[続葉有]

(54) Title: INFORMATION EXTRACTION SYSTEM, INFORMATION EXTRACTION METHOD, INFORMATION EXTRACTION PROGRAM, AND INFORMATION SERVICE SYSTEM

(54) 発明の名称: 情報抽出システム、情報抽出方法、情報抽出プログラムおよび情報サービスシステム



- 100 ... INPUT DEVICE
11 ... PATTERN CANDIDATE CREATING MEANS
12 ... WORD AND PHRASE CANDIDATE CREATING MEANS
13 ... WORD AND PHRASE SELECTING MEANS
200 ... DATA PROCESSOR
300 ... OUTPUT DEVICE
400 ... STORAGE DEVICE
15 ... DOCUMENT GROUP
16 ... PATTERN CANDIDATE
17 ... WORD AND PHRASE CANDIDATES

(57) Abstract: Words and phrases of the same type can be extracted from a plurality of documents of various forms. A storage device (400) stores the documents of various forms. A pattern candidate creating means (11) receives a list of input words selected as samples from words and phrases which are to be included in a dictionary. The pattern candidate creating means (11) selects one document, determines character strings before and after the input word in the document as pattern candidates, and stores them as pattern candidates (16). The pattern candidate creating means (11) performs this processing for each document. A word and phrase candidate creating means (12) extracts the words and phrases sandwiched between the patterns included in the pattern candidate (16) as word and phrase candidates which are to be outputted and stores them as a word and phrase candidate (17). A word and phrase selecting means (13) outputs a word and phrase candidate satisfying a predetermined condition out of the word and phrase candidates included in the word and phrase candidate (17) as an output word to an output device (300).

[続葉有]



MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM,
PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL,
SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ,
VC, VN, ZA, ZM, ZW.

IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK,
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
ML, MR, NE, SN, TD, TG).

- (84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE,

添付公開書類:

— 国際調査報告書

2文字コード及び他の略語については、定期発行される各PCTガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

(57) 要約: 様々な形式の複数の文書から同種の語句を抽出することができるようにする。記憶装置400は、様々な形式の複数の文書を記憶する。パターン候補作成手段11は、辞書に含めるべき語句の中からサンプルとして選択された入力語のリストを受け取る。パターン候補作成手段11は、文書を1つ選択し、その文書における入力語の前後の文字列をパターンの候補として決定し、パターン候補16として記憶させる。パターン候補作成手段11は、この処理を各文書毎に行う。語句候補作成手段12は、パターン候補16に含まれる各パターンに挟まれる語句を出力すべき語句の候補として抽出し、語句候補17として記憶させる。語句選択手段13は、語句候補17に含まれる語句の候補のうち、所定の条件を満足する語句の候補を出力語として出力装置300に出力させる。

明 細 書

情報抽出システム、情報抽出方法、情報抽出プログラムおよび情報サービスシステム

技術分野

- [0001] 本発明は、情報抽出システム、情報抽出方法、情報抽出プログラムおよび情報サービスシステムに関し、特に利用者の所望の種類の語句を文書群より的確に抽出する情報抽出システム、情報抽出方法、情報抽出プログラム、および情報抽出システムを適用した情報サービスシステムに関する。

背景技術

- [0002] ある語句の種類(例えば、人名、地名等の種類)を知るためには、種類に応じた各種辞書(例えば、人名辞書や地名辞書)にその語句が掲載されているかどうかを調べればよい。例えば、ある語句が人名辞書に掲載されていれば、その語句の種類が人名であることがわかる。この場合、ある語句の種類を知るには、その種類に応じた辞書が必要である。
- [0003] ここで、語句とは、単語または複数の単語のまとまりを指すものとする。複数の単語のまとまりの例として、例えば、複数の単語からなるフレーズ、ことわざや慣用句等が挙げられる。また、人名や地名等の固有名詞も語句の概念に含まれる。また、辞書とは、同種の語句のリストであるものとする。
- [0004] このような辞書を作成する方法として、人が様々な大量の文書を読み、多くの語句を語句の種類に応じて分類し、辞書に含まれる語句として登録する方法が挙げられる。この方法では、同種の語句が集まっているという信頼性の高い辞書を作成できるが、人手による作業となるので手間がかかってしまう。
- [0005] また、非特許文献1に辞書作成方法が記載されている。非特許文献1に記載された辞書作成方法では、同じ形式の文書群からパターンを自動的に作成し、パターン間の単語を抽出し、その単語を辞書に登録する。ここで、「同じ形式の文書」とは、文書中で抽出対象となる語句が同じパターンに隣接して出現している文書のことである。また、「パターン」とは、辞書に含める語句(抽出対象となる語句)とそうでない語句と

を区切る文字列である。パターンには、抽出対象となる語句の前に位置するパターン(以下、前方パターンと記す。)と、抽出対象となる語句の後に位置するパターン(以下、後方パターン)とがある。

[0006] 非特許文献1に記載された方法で会社名の辞書を作成する場合の例を以下に示す。まず、会社名が例えば表形式で整列して記述された同じ形式の文書群を、人が収集する。次に、人が、その文書群の中から数個の文書を選択し、その文書に含まれる会社名のリストを作成する。次に、プログラムに従って情報処理装置が、先に選択された文書中に出現する会社名の前方パターンおよび後方パターンを自動的に特定し、前方パターンおよび後方パターンに挟まれた単語(本例では会社名)を抽出する。最後に、抽出した単語を辞書に登録する。このように非特許文献1に記載された方法では、人が、サンプルとして選択した文書とその文書中出现する全ての単語リストを情報処理装置に与えることで、情報処理装置が自動的に辞書を作成する。

[0007] また、特許文献1には、二つの単語a, bのそれぞれの前後に文字列x, yを結合したxay, xbyに関してスコア関数を定義して、二つの単語の関連性を判定することが記載されている。

[0008] 非特許文献1:ニコラス・クシュメリック(Nicholas Kushmerick),「ラッパー・インダクション:エフィシエンシー・アンド・エクスプレシブネス(Wrapper induction: Efficiency and expressiveness)」,アーティフィシャル・インテリジェンス Vol.118(2000)(Artificial Intelligence 118(2000)), 2000年, p. 15-68

特許文献1:特開2003-256447号公報(段落0029-0032)

発明の開示

発明が解決しようとする課題

[0009] 非特許文献1に記載された方法では、同じ形式の文書群が入力されることを前提としている。従って、様々な形式の複数の文書から語句を抽出することは出来なかった。様々な形式の複数の文書から語句を抽出しようとする場合には、それぞれの形式の文書群毎に、人がサンプルとなる文書を選択し、その文書の中に含まれる語句のリストを作成しなければならない。そのため、手間がかかってしまうという問題があった。文書の形式には、例えば、スペースで語句を区切った表形式、タブで語句を区切っ

た表形式、リスト形式、階層形式等様々な種類の形式がある。このような多くの種類の形式毎にサンプルとなる文書の選択や語句のリストの作成を行うことは非特許文献1に記載された方法の利用者にとって負担となる。

[0010] そこで本発明は、様々な形式の複数の文書から同種の語句を抽出することができる情報抽出システム、情報抽出方法、情報抽出プログラム、およびそのような情報抽出システムを適用した情報サービスシステムを提供することを目的とする。

課題を解決するための手段

[0011] 本発明による情報抽出システムは、文書内の語句を文書から抽出して出力する情報抽出システムであって、複数の語句を含む入力語リストが入力される入力手段と、形式が異なる文書を含む複数の文書の中から一つの文書を選択し、選択した文書内で、入力語リストに含まれる語句を抽出対象でない文字列と区切る文字列をパターンとして決定することを文書毎に行うパターン決定手段と、パターンによって区切られる文字列を、パターンの決定に用いた文書から抽出して語句の候補とする語句候補抽出手段と、語句候補抽出手段によって抽出された語句の候補または語句の候補に含まれる部分文字列のうち所定の条件を満たす語句の候補または部分文字列を出力対象の語句として選択する語句選択手段とを備えたことを特徴とする。

[0012] 本発明によれば、パターン決定手段が、形式が異なる文書を含む複数の文書の中から一つの文書を選択し、選択した文書内で、入力語リストに含まれる語句を抽出対象でない文字列と区切る文字列をパターンとして決定することを文書毎に行う。そして、語句候補抽出手段が、パターンによって区切られる文字列を、パターンの決定に用いた文書から抽出して語句の候補とし、語句選択手段が、語句候補抽出手段によって抽出された語句の候補または語句の候補に含まれる部分文字列のうち所定の条件を満たす語句の候補または部分文字列を出力対象の語句として選択する。従って、形式が異なる文書を含む複数の文書の中から同種の語句を抽出することができる。また、語句選択手段が、所定の条件を満たす語句の候補または部分文字列を出力対象の語句として選択するので、出力する語句の信頼性(同種の語句であることの信頼性)を高めることができる。

[0013] パターン決定手段が、入力語リストに含まれる語句の前方の文字列および後方の

文字列をパターンとして決定し、語句候補抽出手段が、前方の文字列および後方の文字列によって挟まれる文字列を文書から抽出して語句の候補とし、語句選択手段が、語句候補抽出手段によって抽出された語句の候補の中から出力対象の語句を選択する構成であつてもよい。

[0014] 語句選択手段が、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを語句の候補毎に算出し、語句スコアが所定のしきい値以上である語句の候補を出力対象の語句として選択する構成であつてもよい。

[0015] 語句選択手段が、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを語句の候補毎に算出し、語句スコアの高い順に語句の候補をソートし所定の順位以上である語句の候補を出力対象の語句として選択する構成であつてもよい。

[0016] 語句選択手段が、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを語句の候補毎に算出し、語句スコアの高い順に語句の候補をソートし上位の所定の割合の語句の候補を出力対象の語句として選択する構成であつてもよい。

[0017] パターン決定手段が、入力語リストに含まれる語句の前方の文字列のみ、あるいは、入力語リストに含まれる語句の後方の文字列のみをパターンとして決定し、語句候補抽出手段が、前方の文字列に続く文字列、あるいは、後方の文字列の直前の文字列を文書から抽出して語句の候補とし、語句選択手段が、語句の候補に含まれる部分文字列を作成し、作成した各部分文字列の中から出力対象の語句を選択する構成であつてもよい。そのような構成によれば、不要な接尾辞あるいは接頭辞を除外した語句を出力することができる。

[0018] 語句選択手段が、語句の候補の総数に対する部分文字列を含む語句の候補の割合を、各語句の候補毎および各部分文字列毎に算出して、前記割合が所定の範囲に属する部分文字列を特定し、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを、特定した部分文字列を含む語句の候補毎に算出して、算出結果を前記部分文字列の語句ス

コアとし、語句スコアが所定のしきい値以上である部分文字列を出力対象の語句として選択する構成であってもよい。

[0019] 語句選択手段が、語句の候補の総数に対する部分文字列を含む語句の候補の割合を、各語句の候補毎および各部分文字列毎に算出して、前記割合が所定の範囲に属する部分文字列を特定し、語句の候補の特定に用いたパターン of 候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを、特定した部分文字列を含む語句の候補毎に算出して、算出結果を前記部分文字列の語句スコアとし、語句スコアの高い順に部分文字列をソートし所定の順位以上である部分文字列を出力対象の語句として選択する構成であってもよい。

[0020] 語句選択手段が、語句の候補の総数に対する部分文字列を含む語句の候補の割合を、各語句の候補毎および各部分文字列毎に算出して、前記割合が所定の範囲に属する部分文字列を特定し、語句の候補の特定に用いたパターン of 候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを、特定した部分文字列を含む語句の候補毎に算出して、算出結果を前記部分文字列の語句スコアとし、語句スコアの高い順に部分文字列をソートし上位の所定の割合の部分文字列を出力対象の語句として選択する構成であってもよい。

[0021] 形式が異なる文書を含む複数の文書であって、予め記憶装置に記憶されている複数の文書の中から、入力語リストから選択した複数の語句全てを含む文書を検索する文書検索手段を備え、パターン決定手段が、文書検索手段によって検索された文書毎にパターンの決定を行う構成であってもよい。そのような構成によれば、同種の語句が出現する可能性の高い文書を対象にしてパターンの決定をおこない、そのような文書においてパターンで区切られる文字列を語句の候補とするので、出力する語句の信頼性を高めることができる。

[0022] 語句選択手段によって選択された語句のリストを入力語リストとして文書検索手段に付与する再作成手段を備えた構成であってもよい。そのような構成によれば、入力語リストを付与された文書検索手段が再度、動作を開始するので、出力対象の語句として選択される語句の数を増加することができる。その結果、同種の語句を網羅的に抽出することができる。

[0023] 語句選択手段によって選択された語句のリストを入力語リストとしてパターン決定手段に付与する再作成手段を備えた構成であってもよい。そのような構成によれば、入力語リストを付与されたパターン決定手段が再度、動作を開始するので、出力対象の語句として選択される語句の数を増加することができる。その結果、同種の語句を網羅的に抽出することができる。

[0024] また、本発明による情報サービスシステムは、上述のいずれかの情報抽出システムと、同種の語句のリストである辞書を記憶する辞書データベースと、辞書の作成および辞書の販売に伴い授受される金額を管理する辞書サービスシステムとを備え、辞書サービスシステムが、入力語リスト作成者によって作成された入力語リストを情報抽出システムに入力して、情報抽出システムから語句のリストを受け取り、語句のリストおよび入力語リストに含まれる語句を含む辞書を辞書データベースに登録する登録手段と、辞書購入者の端末からの要求に応じて、辞書データベースに記憶されている辞書をその端末に送信する辞書提供手段と、辞書作成および辞書の販売に伴い授受される金額を記憶する金銭情報記憶手段と、辞書サービスシステムの運用者が入力語リスト作成者に対し辞書の作成の対価として支払うべき金額、および辞書購入者が辞書サービスシステムの運用者に辞書の対価として支払うべき金額を金銭情報記憶手段に記憶させる金額情報登録手段とを含むことを特徴とする。

[0025] 流行している語句を記憶する流行キーワード記憶手段を備え、辞書サービスシステムが、流行キーワード記憶手段から流行している語句を読み込んで、入力語リスト作成者の端末に送信する流行キーワード提供手段を含む構成であってもよい。

また、本発明による情報サービスシステムは、上述のいずれかの情報抽出システムと、同種の語句のリストである辞書を記憶する辞書データベースと、広告を表示する広告表示装置からキーワードを受信して、広告表示装置に広告を送信する広告サービスシステムとを備え、広告サービスシステムが、入力語リスト作成者によって作成された入力語リストを情報抽出システムに入力して、情報抽出システムから語句のリストを受け取り、語句のリストおよび入力語リストに含まれる語句を含む辞書を辞書データベースに登録する登録手段と、広告および広告に関連する関連キーワードを対応付けて記憶する広告記憶手段と、広告表示装置から受信したキーワードを含む辞書を

辞書データベースに記憶された辞書の中から検索し、検索された辞書に含まれる語句に対応付けられた広告および広告表示装置から受信したキーワードに対応づけられた広告を広告記憶手段から読み込んで、広告表示装置に送信する広告提供手段とを含むことを特徴とする。

[0026] また、本発明による情報抽出方法は、文書内の語句を文書から抽出する情報抽出方法であって、入力手段が、複数の語句を含む入力語リストを入力し、パターン決定手段が、形式が異なる文書を含む複数の文書の中から一つの文書を選択し、選択した文書内で、入力語リストに含まれる語句を抽出対象でない文字列と区切る文字列をパターンとして決定することを文書毎に行い、語句候補抽出手段が、パターンによって区切られる文字列を、パターンの決定に用いた文書から抽出して語句の候補とし、語句選択手段が、語句候補抽出手段によって抽出された語句の候補または語句の候補に含まれる部分文字列のうち所定の条件を満たす語句の候補または部分文字列を出力対象の語句として選択することを特徴とする。

[0027] パターン決定手段が、入力語リストに含まれる語句の前方の文字列および後方の文字列をパターンとして決定し、語句候補抽出手段が、前方の文字列および後方の文字列によって挟まれる文字列を文書から抽出して語句の候補とし、語句選択手段が、語句候補抽出手段によって抽出された語句の候補の中から出力対象の語句を選択する方法であってもよい。

[0028] 語句選択手段が、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを語句の候補毎に算出し、語句スコアが所定のしきい値以上である語句の候補を出力対象の語句として選択する方法であってもよい。

[0029] 語句選択手段が、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを語句の候補毎に算出し、語句スコアの高い順に語句の候補をソートし所定の順位以上である語句の候補を出力対象の語句として選択する方法であってもよい。

[0030] 語句選択手段が、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを語句の候補毎に算出

し、語句スコアの高い順に語句の候補をソートし上位の所定の割合の語句の候補を出力対象の語句として選択する方法であってもよい。

[0031] パターン決定手段が、入力語リストに含まれる語句の前方の文字列のみ、あるいは、入力語リストに含まれる語句の後方の文字列のみをパターンとして決定し、語句候補抽出手段が、前方の文字列に続く文字列、あるいは、後方の文字列の直前の文字列を文書から抽出して語句の候補とし、語句選択手段が、語句の候補に含まれる部分文字列を作成し、作成した各部分文字列の中から出力対象の語句を選択する方法であってもよい。

[0032] 語句選択手段が、語句の候補の総数に対する部分文字列を含む語句の候補の割合を、各語句の候補毎および各部分文字列毎に算出して、前記割合が所定の範囲に属する部分文字列を特定し、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを、特定した部分文字列を含む語句の候補毎に算出して、算出結果を前記部分文字列の語句スコアとし、語句スコアが所定のしきい値以上である部分文字列を出力対象の語句として選択する方法であってもよい。

[0033] 語句選択手段が、語句の候補の総数に対する部分文字列を含む語句の候補の割合を、各語句の候補毎および各部分文字列毎に算出して、前記割合が所定の範囲に属する部分文字列を特定し、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを、特定した部分文字列を含む語句の候補毎に算出して、算出結果を前記部分文字列の語句スコアとし、語句スコアの高い順に部分文字列をソートし所定の順位以上である部分文字列を出力対象の語句として選択する方法であってもよい。

[0034] 語句選択手段が、語句の候補の総数に対する部分文字列を含む語句の候補の割合を、各語句の候補毎および各部分文字列毎に算出して、前記割合が所定の範囲に属する部分文字列を特定し、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを、特定した部分文字列を含む語句の候補毎に算出して、算出結果を前記部分文字列の語句スコアとし、語句スコアの高い順に部分文字列をソートし上位の所定の割合の部分文

字列を出力対象の語句として選択する方法であってもよい。

[0035] 文書検索手段が、形式が異なる文書を含む複数の文書であって、予め記憶装置に記憶されている複数の文書の中から、入力語リストから選択した複数の語句全てを含む文書を検索し、パターン決定手段が、文書検索手段によって検索された文書毎にパターンの決定を行う方法であってもよい。

[0036] 再作成手段が、語句選択手段によって選択された語句のリストを入力語リストとして文書検索手段に付与する方法であってもよい。

[0037] 再作成手段が、語句選択手段によって選択された語句のリストを入力語リストとしてパターン決定手段に付与する方法であってもよい。

[0038] また、本発明による情報抽出プログラムは、文書内の語句を文書から抽出して出力するコンピュータに搭載される情報抽出プログラムであって、コンピュータに、複数の語句を含む入力語リストを入力する入力処理、形式が異なる文書を含む複数の文書の中から一つの文書を選択し、選択した文書内で、入力語リストに含まれる語句を抽出対象でない文字列と区切る文字列をパターンとして決定することを文書毎に行うパターン決定処理、パターンによって区切られる文字列を、パターンの決定に用いた文書から抽出して語句の候補とする語句候補抽出処理、および語句候補抽出処理で抽出された語句の候補または語句の候補に含まれる部分文字列のうち所定の条件を満たす語句の候補または部分文字列を出力対象の語句として選択する語句選択処理を実行させることを特徴とする。

発明の効果

[0039] 本発明によれば、様々な形式の複数の文書から同種の語句を抽出することができる。

図面の簡単な説明

[0040] [図1]本発明による情報抽出システムの第1の実施の形態を示すブロック図である。

[図2]第1の実施の形態におけるデータ処理装置の動作の例を示すフローチャートである。

[図3]文書群の一例を示す説明図である。

[図4]入力語リストの一例を示す説明図である。

[図5]パターン候補の一例を示す説明図である。

[図6]語句候補の一例を示す説明図である。

[図7]出力語リストの一例を示す説明図である。

[図8]第2の実施の形態の動作の例を示すフローチャートである。

[図9]文書群の一例を示す説明図である。

[図10]第2の実施の形態におけるパターン候補の一例を示す説明図である。

[図11]第2の実施の形態における語句候補の一例を示す説明図である。

[図12]本発明による情報抽出システムの第3の実施の形態を示すブロック図である。

[図13]本発明による情報抽出システムの第4の実施の形態を示すブロック図である。

[図14]情報抽出システムの具体的な構成例を示すブロック図である。

[図15]本発明の第5の実施の形態を示すブロック図である。

[図16]辞書サービスシステムの構成例を示すブロック図である。

[図17]本発明の第7の実施の形態を示すブロック図である。

[図18]本発明の第8の実施の形態を示すブロック図である。

[図19]広告サービスシステムと広告閲覧システムの構成例を示すブロック図である。

符号の説明

- [0041]
- 11 パターン候補作成手段
 - 12 語句候補作成手段
 - 13 語句選択手段
 - 14 文書検索手段
 - 15 文書群
 - 16 パターン候補
 - 17 語句候補
 - 18 検索結果文書群
 - 19 再作成手段
 - 100 入力装置
 - 200 データ処理装置
 - 300 出力装置

400 記憶装置

500 情報抽出プログラム

発明を実施するための最良の形態

[0042] 以下、本発明の実施の形態を図面を参照して説明する。

[0043] 実施の形態1.

図1は、本発明による情報抽出システムの第1の実施の形態を示すブロック図である。第1の実施の形態の情報抽出システムは、入力装置100と、プログラム制御により動作するデータ処理装置200と、出力装置300と、情報を記憶する記憶装置400とを含む。

[0044] 入力装置100は、情報が入力される入力デバイスである。入力装置100として、例えばキーボードやマウス等が用いられる。

[0045] 出力装置300は、情報を出力する出力装置である。出力装置300として、例えば、ディスプレイ装置、あるいはプリンタ等が用いられる。

[0046] 記憶装置400は、例えば、ハードディスクドライブ等の記憶装置であり、文書群15とパターン候補16と語句候補17とを記憶する。

[0047] 文書群15は、語句を抽出する対象となる複数の文書である。具体的には、文書群15は、個々の文書を識別するための識別情報(以下、文書IDと記す。)と個々の文書本文とを含む情報である。各文書の本文と各文書IDとは対応づけられている。

[0048] パターン候補16は、パターン候補作成手段11によって決定されたパターンの候補を含む情報である。具体的には、パターン候補16は、パターンの候補、そのパターンの候補の決定に用いた文書の文書ID、およびパターンスコアを含む情報である。パターン候補作成手段11とパターンスコアについては後述する。なお、既に述べたように、「パターン」とは、辞書に含める語句(抽出対象となる語句)とそうでない語句とを区切る文字列であり、パターンには、抽出対象となる語句の前に位置する前方パターンと、抽出対象となる語句の後に位置する後方パターンとがある。本実施の形態では、パターン候補16は、パターンの候補として、前方パターン候補(前方パターンの候補となる文字列)および後方パターン候補(後方パターンの候補となる文字列)を含む。

- [0049] 語句候補17は、語句候補作成手段12によって文書本文から抽出された語句を含む情報である。具体的には、語句候補17は、語句候補作成手段12によって特定された語句の候補、文書ID、パターンスコアを含む情報である。語句候補作成手段12については後述する。なお、既に述べたように、語句とは、単語または複数の単語のまとまりであり、単語だけでなく、複数の単語からなるフレーズ等も語句に該当する。
- [0050] 図1では、文書群15、パターン候補16、および語句候補17が同一の記憶装置400に記憶される場合を示しているが、文書群15、パターン候補16、および語句候補17は複数の記憶装置に分けて記憶されていてもよい。例えば、文書群15、パターン候補16、および語句候補17のうちの一部分が、インターネット等の通信ネットワークを介してデータ処理装置200に接続されている情報処理装置(図示せず。)に記憶されていてもよい。
- [0051] データ処理装置200は、パターン候補作成手段11と、語句候補作成手段12と、語句選択手段13とを含む。
- [0052] パターン候補作成手段11は、入力装置100に入力された入力語を入力データとして、文書群15を参照し、文書内における入力語の出現位置の前後の文字列からパターンの候補を決定する。そして、パターン候補作成手段11は、そのパターンの候補と、その候補の決定に用いた文書の文書IDと、パターンスコアとを、パターン候補16として記憶装置400に記憶させる。
- [0053] 入力語とは、作成しようとする種類の辞書(同種の語句のリスト)に含めるべき語句のうち情報抽出システムの利用者がサンプルとして選択した語句である。
- [0054] パターンスコアとは、パターンまたはパターンの候補の重要度を示すスコアである。パターンスコアは、例えば、文書内に出現する全ての入力語の延べ数に対する、パターンの候補によって区切られる入力語の数の割合で示される。本実施の形態では、文書内に出現する全ての入力語の数に対する、前方パターン候補および後方パターン候補に挟まれた入力語の数の割合をパターンスコアとする。
- [0055] 語句候補作成手段12は、パターン候補16に含まれるデータを入力データとして、文書群15を参照し、パターン候補16においてパターンの候補に対応付けられた文書IDに対応する文書本文を、文書群15から読み込む。そして、語句候補作成手段1

2は、その文書本文におけるパターンの候補（本実施の形態では前方パターン候補および後方パターン候補）の出現位置を特定し、前方パターン候補および後方パターン候補に挟まれる文字列を文書本文から抽出する。ここで抽出された文字列が語句の候補となる。語句候補作成手段12は、語句の候補、語句の候補の抽出に用いたパターンの候補の識別情報（以下、パターンIDと記す。）、およびそのパターンの候補のパターンスコアを対応付けて、語句候補17として記憶装置400に記憶させる。

[0056] 語句選択手段13は、語句候補17を入力データとして、語句の候補のうち、所定の条件を満たす語句を選択し、その語句を出力装置300に出力する。語句選択手段13は、語句の候補の特定に用いたパターンの候補の重要度の高さや、語句の候補が出現する文書の多さを示す数値が高い語句を選択する。

[0057] 次に、動作について説明する。

図2は、本実施の形態におけるデータ処理装置200の動作の例を示すフローチャートである。

[0058] 情報抽出システムは、動作前に文書群15をあらかじめ記憶装置400に記憶しておく。文書群15の一例を図3に示す。文書群15は、文書を識別するための文書IDと文書本文を含み、1レコードは1文書に対応する。1レコードには、1つの文書IDと、その文書IDに対応する文書本文が含まれる。図3に示す例では、「文書A」等の各文書IDおよびその文書IDに対応する文書本文を図示している。文書本文のうち図示を省略した部分は、図3において「…」と表している。

[0059] 図3は、文書群15の一例を示すものであり、ファイルシステムで管理されるファイルや、インターネットを介して入手可能な文書を文書群15として記憶装置400に記憶していてもよい。例えば、文書IDとしてファイルパスやURLを記憶し、文書本文として、そのファイルパスやURLに対応するデータを記憶していてもよい。

[0060] まず、利用者によって、入力装置100（図1参照。）に入力語のリスト（以下、入力語リストと記す。）が入力される。入力語リストは、作成しようとする種類の辞書（同種の語句のリスト）に含めるべき語句のうち情報抽出システムの利用者がサンプルとして選択した数個の語句である。換言すれば、入力語リストは、利用者の所望の種類の数個

の語句である。入力語リストの一例を図4に示す。図4に示す例では、製品名という種類に該当する語句の入力語リストを表している。

[0061] 入力装置100は、入力された入力語リストをデータ処理装置200に送る。すなわち、データ処理装置200には、入力装置100から入力語リストが入力される。入力語リストが入力された後、データ処理装置200は、図2に示すフローチャートに沿って動作する。

[0062] まず、パターン候補作成手段11が文書群15から1レコード分のデータ(1つの文書IDおよびその文書IDに対応する文書本文)を取り出し、その文書本文における入力語の出現位置を特定する(図2に示すステップS1)。パターン候補作成手段11は、ステップS1で、入力語リストに含まれる各入力語の出現位置を特定する。なお、ステップS1で取り出すデータは1レコードである。従って、ステップS1から後述のステップS3までの処理は、1文書毎に行う。

[0063] 例えば、ステップS1で図3に示す文書ID「文書B」に対応する文書本文を選択して取り出しているとする。文書ID「文書B」に対応する文書本文には、図4に示す入力語リストのうち、「製品A」と「製品B」と「製品C」が出現している。パターン候補作成手段11は、選択した文書本文におけるこれらの各入力語の出現位置を特定する。

[0064] 出現位置は、各入力語の文字列の開始位置および終了位置によって表されるが、他の表し方で出現位置を表してもよい。例えば、入力語の文字列の開始位置と入力語の長さの組み合わせ等によって出現位置を表してもよい。本例では、入力語の文字列の開始位置および終了位置で出現位置を表すものとする。

[0065] 次に、パターン候補作成手段11は、文書本文から、入力語の出現位置の前方と後方の文字列を所定の文字数分だけ取得して、パターンの候補を決定する(図2に示すステップS2)。

[0066] パターン候補作成手段11は、以下のようにパターンの候補を決定する。すなわち、パターン候補作成手段11は、取得した各前方の文字列のうち、共通部分を抽出し、その共通部分をパターンの候補(前方パターン候補)として決定する。共通部分は、少なくとも2つの前方の文字列で共通であればよい。ただし、その共通部分は、入力語の直前に位置することを前提とする。また、パターン候補作成手段11は、取得した

所定の文字数分の前方の文字列全体が他の前方の文字列と共通でなくても、その所定の文字数分の前方の文字列全体をパターン候補(前方パターン候補)として決定する。同様に、パターン候補作成手段11は、取得した各後方の文字列のうち、共通部分を抽出し、その共通部分をパターン候補(後方パターン候補)として決定する。共通部分は、少なくとも2つの後方の文字列で共通であればよい。ただし、その共通部分は、入力語の直後であることを前提とする。また、パターン候補作成手段11は、取得した所定の文字数分の後方の文字列全体が他の後方の文字列と共通でなくても、その所定の文字数分の文字列全体をパターン候補(後方パターン候補)として決定する。

[0067] 入力語の出現位置の前方と後方から取り出す文字列の文字数(所定の文字数)は、予め定められていてもよい。あるいは、入力装置100を介して、利用者からパターン候補作成手段11に文字数が入力され、パターン候補作成手段11は、その文字数を所定の文字数として定めてもよい。

[0068] 所定の文字数が「10」であるものとして、パターン候補作成手段11が前方パターン候補および後方パターン候補を決定する例を示す。また、ステップS1では、文書ID「文書B」に対応する文書本文(図3参照。)が選択されたものとする。この文書本文では、入力語「製品A」の前方の文字列(10文字分の文字列)は、「size="10">」である。「製品B」の前方の文字列(10文字分)も同様に、「size="10">」である。また、この文書本文では、入力語「製品C」は二回出現している。そのうちの1つの「製品C」の前方の文字列(10文字分)は、「size="10">」であり、もう1つの「製品C」の前方の文字列(10文字分)は、「nt<p>b>・」である。パターン候補作成手段11は、各前方文字列を比較して、共通部分となる文字列を抽出する。本例では、4回出現した入力語のうち3つの入力語で「size="10">」が共通であるので、パターン候補作成手段11は、「size="10">」を前方パターン候補として決定する。また、10文字分の文字列「nt<p>b>・」全体は、「size="10">」と共通ではないが、この10文字分の文字列「nt<p>b>・」も前方パターン候補として決定する。

[0069] 後方パターン候補についても同様に決定する。この文書本文では、入力語「製品A」の後方の文字列(10文字分)は、「
」である。「製品B」の後方の文

字列(10文字分)は、「
」である。2回出現する「製品C」のうち、1つの「製品C」の後方の文字列(10文字分)は、「<p>」であり、もう1つの「製品C」の後方の文字列(10文字分)は、「のカテゴリ」である。パターン候補作成手段11は、各後方文字列を比較して、共通部分となる文字列を抽出する。本例では、4回出現した入力語のうち3つの入力語で「<」が共通であるので「<」を後方パターン候補として決定する。また、4回出現した入力語のうち2つの入力語で「
」が共通であるので、「
」を後方パターン候補として決定する。また、10文字分の文字列「<p>」全体や、「のカテゴリ全体は、他と共通ではないが、パターン候補作成手段11は、「<p>」全体や、「のカテゴリ全体も後方パターン候補として決定する。

[0070] 次に、パターン候補作成手段11は、前方パターン候補群と後方パターン候補群の組合せから、パターンスコアを計算し、その計算結果を記憶装置400に記憶させる(図2に示すステップS3)。ステップS3では、パターン候補作成手段11は、1つの前方パターン候補と1つの後方パターン候補のペアと、そのペアを識別するパターンIDと、パターン候補の決定に用いた文書本文(ステップS1で取り出した文書本文)の文書IDと、パターンスコアとを対応付けて、パターン候補16として記憶装置400に記憶させる。

[0071] 既に説明したように、パターンスコアとは、パターンまたはパターンの候補の重要度を示すスコアである。本実施の形態では、前方パターン候補と後方パターン候補のペア毎にパターンスコアを計算する。パターンスコアとして、例えば、前方パターン候補と後方パターン候補のペアに挟まれた語句が入力語である確率や、パターンの文字列の長さ等が考えられる。本実施の形態では、文書内に出現する全ての入力語の延べ数に対する、前方パターン候補および後方パターン候補に挟まれた入力語の数の割合をパターンスコアとする。すなわち、パターン候補作成手段11は、「前方パターン候補および後方パターン候補に挟まれた入力語の数／文書内に出現する全ての入力語の数」を計算してパターンスコアを算出すればよい。このパターンスコアは、前方パターン候補と後方パターン候補のペアに挟まれた語句が入力語である確

率を表している。前方パターン候補と後方パターン候補のペアによって抽出できる入力語の数が多いほど、そのペアは重要なパターンと考えられるので、本実施の形態では、上記のようにパターンスコアを算出する。

[0072] 先に例示した前方パターン候補および後方パターン候補を用いて、パターンスコアの計算例を示す。例えば、前方パターン候補「size="10">」と後方パターン候補「<」のペアでは、間に挟まれる入力語として「製品A」「製品B」「製品C」がある(図3に示す「文書B」参照。)。従って、前方パターン候補および後方パターン候補に挟まれた入力語の数は3である。また、「製品C」は文書中に2回出現しているので、「文書B」に出現する全ての入力語の数は4である。従って、パターン候補作成手段11は、パターンスコアを「3/4」と算出する。

[0073] パターン候補作成手段11は、他のペアについても同様にパターンスコアを算出する。前方パターン候補「size="10">」と後方パターン候補「
」のペアでは、パターンスコアは「2/4」と算出される。前方パターン候補「size="10">」と後方パターン候補「のカテゴリ」のペアでは、パターンスコアは「0/4」と算出される。

[0074] また、パターン候補作成手段11は、前方パターン候補と後方パターン候補のペアに対してパターンIDを割り当て、パターンID、文書ID、前方パターン候補と後方パターン候補のペア、およびパターンスコアとを対応付けて、パターン候補16として記憶装置400に記憶させる。図5は、パターン候補16の一例を示す説明図である。図5に示すように、パターン候補16は、パターンIDと、文書IDと、前方パターン後方と、後方パターン候補と、パターンスコアとを含んでいる。1レコードは、1組の前方パターン候補と後方パターン候補のペアに対応する。図5に示す例において、図示を省略したレコードは「…」と表している。また、ここでは、計算したパターンスコアが「0」である場合、パターン候補作成手段11がそのパターンスコアを含むレコードを記憶装置400に記憶させない場合を例に示している。図5においても、パターンスコアが「0」となるレコードは示していない。

[0075] ステップS3の後、パターン候補作成手段11は、文書群15に含まれる全ての文書を取り出して、全ての文書に対してステップS3までの処理を実行したか否かを判定

する(図2に示すステップS4)。ステップS1～S3の処理を実行していない文書が残っている場合(ステップS4のN)、ステップS1に移行し、ステップS1以降の処理を繰り返す。全ての文書に対してステップS3までの処理を実行したならば、ステップS5に移行する。

[0076] ステップS5では、語句候補作成手段12は、記憶装置400に記憶されるパターン候補16から1レコードを取り出し、文書IDに対応する文書本文における前方パターン候補と後方パターン候補の出現位置を特定する(図2に示すステップS5)。まず、ステップS5において、語句候補作成手段12は、まず、パターン候補16から1レコードを取り出す。そして、語句候補作成手段12は、そのレコードに含まれる文書IDに対応する文書本文を文書群15の中から読み込む。語句候補作成手段12は、その文書本文の中で、取り出したレコードに含まれる前方パターン候補と後方パターン候補の出現位置を特定する。

[0077] 図5に例示するパターン候補16の2レコード目を語句候補作成手段12が取り出した場合を例にして、ステップS5における語句候補作成手段12の動作を説明する。図5に例示する2レコード目のデータには、文書IDとして「文書B」が含まれている。また、前方パターン候補として「size="10">」が含まれ、後方パターン候補として「<」が含まれている。従って、語句候補作成手段12は、記憶装置400に記憶される文書群15の中から「文書B」に対応する文書本文を読み込む。そして、語句候補作成手段12は、その文書本文における前方パターン候補「size="10">」の出現位置と、後方パターン候補「<」の出現位置を特定する。

[0078] ステップS5の後、語句候補作成手段12は、前方パターン候補の出現位置と後方パターン候補の出現位置に基づいて、その前方パターン候補と後方パターン候補に挟まれる文字列を語句の候補として抽出する。そして、語句候補作成手段12は、抽出した語句の候補と、その語句の候補を抽出した文書の文書IDと、その語句の候補の抽出に用いたパターンの候補(前方パターン候補と後方パターン候補のペア)のパターンIDと、そのパターンスコアとを対応させて、語句候補17として記憶装置400に記憶させる(図2に示すステップS6)。

[0079] ただし、「語句の候補となる文字列の出現位置は、前方パターン候補の出現位置と

後方パターン候補の出現位置を含まない。」という条件を満足するように、語句候補作成手段12は、語句の候補を文書本文から抽出する。すなわち、文書本文から抽出される語句の候補の文字列には、前方パターン候補や後方パターン候補は含まれない。例えば、図5に例示するパターン候補16の2レコード目に含まれる文書ID(「文書B」)に対応する文書本文では、前方パターン候補「size=“10”>」と後方パターン候補「<」に挟まれる文字列として、「製品B
製品A」という文字列が存在する。この文字列の出現位置には、後方パターン候補「<」の出現位置も含まれている。従って、語句候補作成手段12は、このような文字列は語句の候補として採用しない。

[0080] また、語句候補作成手段12は、入力語と同一の文字列も語句の候補として採用しない。従って、入力語と同一の文字列が語句候補17として記憶されることはない。

[0081] 語句候補作成手段12は、このように抽出した語句の候補と、その語句の候補を抽出した文書の文書IDと、その語句の候補の抽出に用いたパターンの候補(前方パターン候補と後方パターン候補のペア)のパターンIDと、そのパターンスコアとを対応させて、語句候補17として記憶装置400に記憶させる。図6は、語句候補17の一例を示す説明図である。図6に示すように、語句候補17は、文書IDと、文書本文から抽出された文字列(語句の候補)と、パターンIDと、パターンスコアとを含んでいる。語句候補17における1レコードは、1つの語句の候補に対応する。

[0082] ステップS6の後、語句候補作成手段12は、パターン候補16に含まれる全てのレコード(全てのパターン候補)についてステップS5、S6の処理を実行したか否かを判定する(図2に示すステップS7)。ステップS5、S6の処理を実行していないレコードがパターン候補16に残っている場合(ステップS7のN)、ステップS5に移行し、ステップS5以降の処理を繰り返す。パターン候補16の全てのレコードに対してステップS5、S6の処理を実行したならば(ステップS7のY)、ステップS8に移行する。

[0083] ステップS8では、語句選択手段13は、語句候補17を参照して、文書から抽出された語句の候補を選択し、その語句の候補の語句スコアを計算する(図2に示すステップS8)。語句選択手段13は、ステップS8において、まず、語句の候補の文字列が同一であるレコードを語句候補17から抜き出し、抜き出したレコードに共通の語句の候

補の語句スコアを計算する。語句スコアとは、語句の候補の特定に用いたパターンの候補の重要度の高さや、語句の候補が出現する文書の多さを示す数値である。語句スコアの値が高いほど、パターンの候補の重要度が高く、語句の候補が出現する文書の数が多いことを示す。語句スコアとして、語句の候補の文字列が同一である各レコードに含まれるパターンスコアの平均値または合計値を用いることができる。また、語句の候補の文字列が同一である各レコードに含まれる文書IDの種類数を語句スコアとして用いてもよい。これらは語句スコアの例示であり、他の計算値を語句スコアとしてもよい。ここでは、語句選択手段13が、語句の候補の文字列が同一であるレコードを語句候補17から抜き出したのち、その各レコードを文書IDで分類して文書ID毎にパターンスコアの平均値を計算し、文書ID毎に計算した平均値の合計を語句スコアとする場合を例にして説明する。この語句スコアの計算方法では、1つの文書にある語句の候補が複数回出現する場合を考慮して、文書ID毎にパターンスコアの平均値を計算している。そして、複数の文書で出現する語句の候補に、より大きなスコアを割り当てるために、文書ID毎に計算したパターンスコアの平均値の合計を語句スコアとしている。

[0084] 例えば、図6の語句候補17に含まれる語句の候補「製品D」の語句スコアは、以下のように計算される。語句の候補が「製品D」であり、文書IDが「文書A」であるレコードは1つしかなく、そのパターンスコアは「3/3」である。語句の候補が「製品D」であり、文書IDが「文書B」であるレコードは2つあり、そのパターンスコアは、それぞれ「3/4」、「2/4」である。この平均値は「 $(3/4 + 2/4) / 2$ 」である。語句の候補が「製品D」であり、文書IDが「文書D」であるレコードは1つしかなく、そのパターンスコアは「2/2」である。従って、語句選択手段13は、「 $3/3 + (3/4 + 2/4) / 2 + 2/2 = 2.625$ 」と語句スコアを計算する。

[0085] 次に、語句候補17の全てのレコードについて、ステップS8の処理を完了したか否かを判定する(図2に示すステップS9)。ステップS8の処理が行われていないレコードが残っている場合、すなわちステップS8で選択されていないレコードが残っている場合(ステップS9のN)、ステップS8に移行し、ステップS8以降の処理を繰り返す。語句候補17の全てのレコードについて、ステップS8の処理を完了した場合(ステップ

S9のY)、ステップS10に移行する。

- [0086] ステップS10では、語句選択手段13は、語句スコアに基づいて、出力すべき語句の候補を選択する(図2に示すステップS10)。出力すべき語句の候補の選択方法として、語句スコアの値が所定のしきい値以上である語句の候補を選択する方法がある。また、語句スコアの高い順に語句の候補をソートし、所定の順位以上である語句の候補を選択する方法がある。また、語句スコアの高い順に語句の候補をソートし、上位の所定の割合の語句の候補を選択してもよい。ここでは、語句スコアの高い順に語句の候補をソートし、上位の所定の割合の語句の候補を選択する場合を例にする。この場合、語句選択手段13は、例えば、語句スコアの高い上位10%の語句の候補を選択する。ここでは、所定の割合として上位10%を例示したが、所定の割合は10%でなくてもよい。また、上記の各選択方法における所定のしきい値、所定の順位、所定の割合は、予め定められていてもよい。あるいは、入力装置100を介して、利用者から入力されてもよい。
- [0087] このように、語句スコアの高い語句の候補を選択することによって、より多くの文書に出現する語句や、重要度の高いパターンによって抽出された語句を選択することができる。
- [0088] なお、既に説明したように、語句の候補の文字列が同一である各レコードに含まれるパターンスコアの平均値あるいは合計値を語句スコアとしてもよい。あるいは、語句の候補の文字列が同一である各レコードに含まれるパターンIDによって特定されるパターンの候補(本実施の形態では前方パターン候補および後方パターン候補)の文字数の平均値を語句スコアとしてもよい。このように語句スコアを計算した場合であっても、出力すべき語句の候補の選択は、上述のように行えばよい。例えば、語句スコアの値が所定のしきい値以上である語句の候補を選択すればよい。あるいは、語句スコアの高い順に語句の候補をソートし、所定の順位以上である語句の候補を選択してもよい。また、あるいは、語句スコアの高い順に語句の候補をソートし、上位の所定の割合の語句の候補を選択してもよい。
- [0089] 最後に、語句選択手段13は、選択した語句の候補の集合を出力語リストとして出力装置300に出力(例えば、表示出力、あるいは印字出力)させる。出力語は、ステッ

プS10で選択された語句の候補である。出力語リストの一例を図7に示す。図7では、選択した語句の候補とともに、その語句スコアも出力する場合を示している。

[0090] 情報抽出システムの利用者は、出力語リストに含まれる出力語と自身が入力した入力語とを含む辞書を作成すればよい。また、利用者は、各出力語に対応する語句スコアを参照して、個々の出力語を辞書に含めるべきか否かを精査してもよい。

[0091] 次に、本実施の形態の効果について説明する。

[0092] 本実施の形態では、パターン候補作成手段11が1つずつ文書を文書群15から取り出し、各文書本文からパターンの候補を決定し、パターンの候補を用いて語句の候補を文書から抽出する。従って、文書形式に依存せずに語句の抽出を行うことができる。すなわち、文書群15に様々な文書形式の文書が含まれていたとしても、各文書から入力語と同種の語句を抽出することができる。

[0093] また、本実施の形態では、さらに、パターン候補作成手段11によって決定されたパターンの候補を用いて、語句候補作成手段12が語句の候補を抽出し、語句選択手段13が語句スコアに基づいて語句の候補を選択する。そして、その選択された語句の候補が出力語として出力される。従って、出力された語句の信頼度を確保することができる。

[0094] 実施の形態2.

第2の実施の形態の情報抽出システムは、第1の実施の形態と同様であり、第2の実施の形態の情報抽出システムも図1のように示すことができる。ただし、パターン候補作成手段11、語句候補作成手段12、語句選択手段13の処理内容が第1の実施の形態とは異なる。

[0095] 第2の実施の形態では、パターン候補作成手段11は、前方または後方のみのパターンの候補を決定する。すなわち、本実施の形態では、パターン候補作成手段11は、パターンの候補として、前方パターン候補のみ、あるいは後方パターン候補のみを決定する。

[0096] また、第2の実施の形態では、語句候補作成手段12は、前方パターン候補のみ、あるいは後方パターン候補のみに基づいて語句候補17を作成する。従って、前方パターン候補のみを考慮した処理、または、後方パターン候補のみを考慮した処理を

実現することができる。

- [0097] 第2の実施の形態では、前方パターン候補のみ、あるいは後方パターン候補のみに基づいて語句の候補となる文字列を決定することで、不必要な接尾辞や接頭辞を除去することが可能となる。不必要な接頭辞や接尾辞とは、文書テキスト中で注意を喚起するために付与する、「*」や「#」といった記号等のことである。
- [0098] 第2の実施の形態では、前方パターン候補のみ、または、後方パターン候補のみを用いることによって、このような不必要な接尾辞や接頭辞の除去を実現し、これによって出力語の信頼性向上を図る。
- [0099] 次に、本実施の形態の動作について説明する。図8は、本実施の形態の動作の例を示すフローチャートである。図8に示すフローチャートは、第1の実施の形態の動作を示すフローチャート(図2参照。)と以下の点で異なる。すなわち、図2に示すステップS2, S3, S5, S6, S8の処理が、それぞれ、ステップS21, S31, S51, S61, S81の処理に置き換わっている点で異なる。ステップS21, S31, S51, S61, S81の処理については後述する。
- [0100] 以下の説明では、第1の実施の形態と異なる処理内容について述べ、第1の実施の形態と同様の処理については説明を省略する。また、以下の例では、パターンの候補として前方パターン候補のみを用いる例を示すが、パターンの候補として後方パターン候補のみを用いる場合も同様である。パターンの候補として前方パターン候補と後方パターン候補のどちらを用いるかを示す情報をデータ処理装置200が保持しておき、その情報に従って前方パターン候補のみ、あるいは後方パターン候補のみを用いてもよい。あるいは、パターンの候補として前方パターン候補か後方パターン候補のどちらを用いるかを示す情報が、入力装置100を介して利用者から入力され、データ処理装置200は、その情報に従って、前方パターン候補のみ、あるいは後方パターン候補のみを用いてもよい。
- [0101] ステップS1の後のステップS21では、パターン候補作成手段11は、前方パターン候補のみ、あるいは後方パターン候補のみを決定する。前方パターン候補を決定する動作、後方パターン候補を決定する動作は、第1の実施の形態と同様である。
- [0102] 例えば、記憶装置400(図1参照。)が文書群15として、図9に例示する文書群を記

憶していて、ステップS1では、文書ID「文書B」に対応する文書本文(図3参照。)が選択されたものとする。なお、図9は、文書群15の一例を示す説明図であり、「文書B」に対応する文書本文において、文字列「製品D」の後に文字「*」が追加されている点で、図3に例示する文書群とは異なっている。また、入力語リストとして、第1の実施の形態で例示した場合と同様に図4に例示する入力語リストが入力されているものとする。ここでは、パターン候補作成手段11が前方パターン候補のみを決定する場合を例示する。前方パターン候補を決定する動作は第1の実施の形態と同様である。従って、パターン候補作成手段11は、「文書B」に対応する文書本文から「size=“10”>」、「nt><p>・」を抽出して、この2つの文字列を前方パターン候補として決定する。

[0103] ステップS21の後、パターン候補作成手段11は、パターンスコアを計算し、その計算結果を記憶装置400に記憶させる(ステップS31)。第2の実施の形態では、パターン候補作成手段11は、前方パターン候補のみ、あるいは後方パターン候補のみからパターンスコアを計算する。本実施の形態では、前方パターン候補のみからパターンスコアを計算する場合、文書内に出現する全ての入力語の延べ数に対する、入力語出現位置の前方の前方パターン候補出現数の割合をパターンスコアとすればよい。すなわち、パターン候補作成手段11は、「入力語出現位置前方の前方パターン候補出現数／文書中に出現する全ての入力語数」を計算してパターンスコアを算出すればよい。また、後方パターン候補のみからパターンスコアを計算する場合、文書内に出現する全ての入力語の延べ数に対する、入力語出現位置の後方の後方パターン候補出現数の割合をパターン候補とすればよい。すなわち、パターン候補作成手段11は、「入力語出現位置後方の後方パターン候補出現数／文書中に出現する全ての入力語数」を計算してパターンスコアを算出すればよい。

[0104] ここでは、前方パターン候補のみからパターンスコアを計算する場合を例示する。例えば、文書ID「文書B」に対応する文書本文を参照すると、「製品A」、「製品B」、「製品C」の入力語の前方における前方パターン候補「size=“10”>」の出現数は3である。また、この文書本文では、入力語「製品A」、「製品B」、「製品C」は、延べ4つ出現している。従って、パターン候補作成手段11は、「size=“10”>」のパターンスコア

コアを「3/4」と算出する。パターン候補作成手段11は、他の前方パターン候補についても同様にパターンスコアを算出する。

[0105] パターン候補作成手段11は、各前方パターン候補(後方パターン候補のみを用いて処理を実行する場合には各後方パターン候補)に対してパターンIDを割り当て、パターンIDと、前方パターン候補(または後方パターン候補)と、その前方パターン候補(または後方パターン候補)の決定に用いた文書本文の文書IDと、パターンスコアとを対応付けて、パターン候補16として記憶装置400に記憶させる。図10は、第2の実施の形態におけるパターン候補16の一例を示す説明図である。図10に示すように、本実施の形態におけるパターン候補16は、パターンIDと、文書IDと、前方パターン候補と、パターンスコアとを含んでいる。1レコードは、1つの前方パターン候補に対応する。ステップS21で後方パターン候補のみを決定した場合には、前方パターン候補の代わりに後方パターン候補を含む。

[0106] ステップS31の後、第1の実施の形態と同様にステップS4の判定処理を行い、文書群15に含まれる全ての文書に対してステップS1, S21, S31の処理を実行したならば、ステップS51に移行する。

[0107] ステップS51では、語句候補作成手段12は、記憶装置400に記憶されるパターン候補16から1レコードを取り出し、文書IDに対応する文書本文における前方パターン候補の出現位置を特定する(ステップS51)。ただし、後方パターン候補のみを用いて処理を実行する場合(すなわちステップS21で後方パターン候補のみを決定した場合)には各後方パターン候補の出現位置を特定する。

[0108] ステップS51の後、語句候補作成手段12は、語句の候補を特定する(ステップS61)。ステップS21で前方パターン候補のみを決定した場合、前方パターン候補に続く所定の文字数の文字列を語句の候補として特定する。また、ステップS21で後方パターン候補のみを決定した場合、後方パターン候補の直前の所定の文字数の文字列を語句の候補として特定する。なお、この所定の文字数は、予め定められていてもよい。あるいは、入力装置100を介して利用者から文字数が入力され、語句候補作成手段12は、その文字数を所定の文字数として定めてもよい。

[0109] 例えば、所定の文字数が6であり、図10に示すパターン候補16の2レコード目に含

まれる前方パターン候補を用いて語句の候補を特定する場合を例示する。図10に示すパターン候補16の2レコード目に含まれる文書IDは、「文書B」である。従って、語句候補作成手段12は、「文書B」の文書本文における前方パターン候補「size＝“10”>」に続く6文字分の文字列を語句の候補を抽出し、語句の候補とする。すなわち、語句候補作成手段12は、「文書B」の文書本文から「製品D*＜／」を抽出し、語句の候補とする。

[0110] ただし、語句候補作成手段12は、前方パターン候補の直後（または、後方パターン候補の直前）に入力語を含む文字列は語句の候補として採用しない。

[0111] 語句候補作成手段12は、特定した語句の候補と、その語句の候補を抽出した文書の文書IDと、その語句の特定に用いたパターンの候補（前方パターン候補あるいは後方パターン候補）のパターンIDと、そのパターンスコアとを対応させて、語句候補17として記憶装置400に記憶させる。図11は、本実施の形態における語句候補17の一例を示す。本実施の形態における語句候補17は、第1の実施の形態の場合と同様に、文書IDと語句の候補とパターンIDとパターンスコアとを含んでいる。語句候補17における1レコードは、1つの語句の候補に対応する。ただし、本実施の形態では、それぞれの語句の候補の文字数は所定の文字数（本例では6）であり、不必要な接尾辞やタグ文字列等が付加されたままになっている。

[0112] ステップ61の後、第1の実施の形態と同様にステップS7の判定処理を行い、パターン候補16（図10参照。）の全てのレコードに対してステップS51、S61の処理を実行したならば、ステップS81に移行する。

[0113] ステップS81では、語句選択手段13は、語句候補17を参照して語句の候補を選択し、その語句の候補の語句スコアを計算する（ステップS81）。ステップS81において、まず、語句選択手段13は、語句候補17から1レコードを抜き出す。語句選択手段13は、そのレコードに含まれる語句の候補の部分文字列を作成する。前方パターン候補から決定された語句の候補の場合、その語句の候補の先頭文字のみからなる部分文字列、先頭文字からその次の文字までで構成した部分文字列、・・・、先頭文字から最終文字までで構成した部分文字列をそれぞれ作成する。後方パターン候補から決定された語句の候補の場合、その語句の候補の最終文字のみからなる部分

文字列、最終文字からその1つ前までの文字までで構成した部分文字列、・・・、最終文字から先頭文字までで構成した部分文字列をそれぞれ作成する。次に、語句選択手段13は、語句候補17を参照し、作成した部分文字列を含むレコードを検索する。そして、語句選択手段13は、部分文字列を含むレコード数の全レコード数に占める割合が所定の範囲内であるか否かを判定する。語句選択手段13は、作成したそれぞれの部分文字列に対してこの判定処理を行い、部分文字列を含むレコード数の全レコード数に占める割合が所定の範囲内となっている部分文字列を語句スコアの計算対象とする。ここで、上記の所定の範囲を示す情報を語句候補選択手段13が予め保持していてもよい。あるいは、所定の範囲を示す情報が入力装置100を介して利用者から入力され、語句選択手段13は、その所定の範囲を用いて上記の判定処理を行ってもよい。所定の割合は、例えば、20%～40%の範囲とすることが好ましいが、この範囲に限定されるわけではない。

[0114] 語句選択手段13は、語句スコアの判定対象とした部分文字列を含むレコードを語句候補17(図11参照。)から抜き出し、その部分文字列の語句スコアを計算する。この語句スコアの計算は第1の実施の形態で説明した計算と同様に行えばよい。

[0115] 図11に例示する語句候補17の1レコード目を取り出した場合を例にしてステップS81の具体例を説明する。この場合、語句選択手段13は、語句の候補「製品D</t」から部分文字列を作成し、「製」、「製品」、「製品D」、「製品D<」、「製品D</」、「製品D</t」の6つの部分文字列を得る。語句選択手段13は、各部分文字列を含むレコードを検索する。部分文字列「製」および「製品」に関しては、図11に示す語句候補17のうち1, 2, 3, 4, 5, 8, 9, 10レコード目の8個のレコードを検索する。部分文字列「製品D」に関しては、1, 4, 9レコード目の3個のレコードを検索する。部分文字列「製品D<」、「製品D</」、「製品D</t」に関しては、それぞれ4レコード目の1個のレコードを検索する。そして、語句選択手段13は、部分文字列を含むレコード数の全レコード数に占める割合が所定の範囲内(本例では20%～40%の範囲内とする。)となっている部分文字列を語句スコアの計算対象とする。図11に示す例では全レコード数は10個なので、「製品D」を語句スコアの計算対象とする。

[0116] 語句選択手段13は、部分文字列「製品D」を含むレコード(1, 4, 9レコード目の3

つのレコード)を語句候補17から抜き出し、語句スコアを計算する。ここでは、第1の実施の形態で示した具体例と同様に、各レコードを文書IDで分類して文書ID毎にパターンスコアの平均値を計算し、文書ID毎に計算した平均値の合計を語句スコアとする場合の例を示す。語句選択手段13は、「 $3/3 + 3/4 + 2/2 = 2.75$ 」と語句スコアを計算する。

- [0117] 次のステップS9では、語句候補17に含まれる全てのレコードを1つずつ抜き出してステップS81の処理を完了したか否かを判定する。まだ、抜き出していないレコードが存在するならば(ステップS9のN)、ステップS81に移行し、ステップS81以降の処理を繰り返す。全てのレコードについてステップS81の処理を完了したならば(ステップS9のY)、ステップS10に移行する。
- [0118] ステップS10では、語句スコアの計算対象とされた文字列(語句の候補の部分文字列)の中から、語句を選択する(ステップS10)。ステップS10における語句の選択方法は、第1の実施の形態と同様である。例えば、語句スコアの値が所定のしきい値以上である部分文字列を選択してもよい。また、例えば、語句スコアの高い順に部分文字列をソートし、所定の順位以上である部分文字列を選択してもよい。また、例えば、語句スコアの高い順に部分文字列をソートし、上位の所定の割合の部分文字列(例えば上位10%の部分文字列)を選択してもよい。上記の各選択方法における所定のしきい値、所定の順位、所定の割合は、予め定められていてもよい。あるいは、入力装置100を介して、利用者から入力されてもよい。
- [0119] 最後に、語句選択手段13は、第1の実施の形態と同様に出力語リストを出力装置300に出力(例えば、表示出力、あるいは印字出力)させる。
- [0120] 情報抽出システムの利用者は、出力語リストに含まれる出力語と自身が入力した入力語とを含む辞書を作成すればよい。また、利用者は、各出力語に対応する語句スコアを参照して、個々の出力語を辞書に含めるべきか否かを精査してもよい。
- [0121] 上述の具体例では、ステップS21で前方パターン候補のみを決定し、前方パターン候補を用いる場合を説明した。ステップS21で後方パターン候補のみを決定し、後方パターン候補を用いる場合の動作も同様である。
- [0122] 次に、第2の実施の形態の効果について説明する。第2の実施の形態では、パター

ンの候補として、前方パターン候補のみ、または後方パターン候補のみを用いる。従って、出力語となる文字列の前あるいは後ろに余計な文字列(例えば、不必要な接尾辞や接頭辞)が付加されていてもこれを除去できる。また、第1の実施の形態と同様に、文書形式に依存せずに語句の抽出を行うことができる。すなわち、文書群15に様々な文書形式の文書が含まれていたとしても、各文書から入力語と同種の語句を抽出することができる。また、出力された語句の信頼度を確保することができる。

[0123] 実施の形態3.

図12は、本発明による情報抽出システムの第3の実施の形態を示すブロック図である。第1の実施の形態と同様の構成部については、図1と同一の符号を付し、説明を省略する。第3の実施の形態では、データ処理装置200は、パターン候補作成手段11、語句候補作成手段12、語句選択手段13に加え、文書検索手段14も含む。また、記憶装置400は、文書群15、パターン候補16、語句候補17に加え、検索結果文書群18も記憶する。

[0124] 文書検索手段14は、入力装置100から入力語リストを受け取り、入力語リストの中から複数の入力語を選択し、選択した入力語群を含む文書本文およびその文書IDを文書群15の中から検索する。文書検索手段14は検索した文書本文および文書IDの組を検索結果文書群18として記憶装置400に記憶させる。

[0125] 検索結果文書群18は、文書検索手段14によって文書群15から検索された検索結果である。検索結果文書群18は、文書群15と同様に、文書IDと文書本文とを含む。パターン候補作成手段11は、検索結果文書群18から各レコードを読み込んで、パターン候補16を作成する。また、語句候補作成手段12は、検索結果文書群18から各レコードを読み込んで、語句候補17を作成する。

[0126] このように構成によって、文書内での同種語句のまとまりやすさを考慮して、語句抽出対象となる文書を効果的に文書群15の中から選択することができる。そして、その後の処理では、文書群15ではなく、文書群15から選択された文書を用いて処理を行うので、同種の語句を含んでいる信頼性がより高い出力語リストを作成することができる。

[0127] 一般的に同じ種類の語句は同じ文書に出現することが多い。例えば、語句の種類

として「会社名」を例にして説明する。会社名が複数出現する文書は、会社名が1つしか出現しない文書に比べて、リスト形式や表形式など同じようなパターンの間に会社名が出現することが多いと考えられる。したがって、このような文書のみを対象にしてパターン候補16を作成し、語句候補17を作成することで、会社名以外の語句を抽出してしまう可能性を下げることができる。そこで、同一文書に複数の入力語が出現する文書をあらかじめ検索することで、同様の語句が出現する文書のみを対象に信頼性の高い出力語リストを作成できる。

[0128] 第3の実施の形態では、文書検索手段14が、入力語リストから複数の入力語を選択し、文書群15に含まれる文書のうち、選択した入力語を全て含む文書を検索する。そして、パターン候補作成手段11および語句候補作成手段12は、検索された文書(すなわち、検索結果文書群18)を用いて処理を行う。この結果、出力語リストの信頼性を向上させることができる。

[0129] 次に、本実施の形態における処理経過の例について説明する。第1の実施の形態と同様に、情報抽出システムは、動作前に文書群15をあらかじめ記憶装置400に記憶しておく。また、利用者によって、入力装置100に入力語リストが入力される。入力装置100は、入力された入力語リストをデータ処理装置200に送る。すなわち、データ処理装置200には、入力装置100から入力語リストが入力される。

[0130] データ処理装置200の文書検索手段14は、入力装置100から入力語リストが送られてくると、その入力語リストの中から複数の入力語を選択する。文書検索手段14が入力語リストの中から選択する入力語の数は、文書検索手段14が予め記憶していてもよい。あるいは、選択すべき入力語の数が、入力装置100を介して利用者から文書検索手段14に入力されてもよい。文書検索手段14が入力語リストの中から選択する入力語の数は、例えば3とすることが好ましいが、3に限定されるわけではない。ただし、選択する入力語の数は2以上として、入力語リストから複数の入力語が選択されるように定める。

[0131] 文書検索手段14は、定められた数の入力語を選択するときに、入力語リストの中から定められた数の入力語をランダムに選択することが好ましい。ただし、このような選択方法に限定されるわけではなく、例えば、文書群15での出現回数頻度が多い入

力語から順に選択してもよい。

- [0132] 次に、文書検索手段14は、文書群15を参照し、選択した入力語群(複数の入力語)を全て含む文書を検索し、検索結果文書群18として記憶装置400に記憶させる。すなわち、文書検索手段14は、選択した複数の入力語を全て含む文書本文およびその文書IDを文書群15から検索し、検索した文書IDおよび文書本文を検索結果文書群18として記憶装置400に記憶させる。
- [0133] 次に、文書検索手段14が、十分な検索結果文書群18が得られたか否かを判定する。十分な検索結果文書群18が得られていないと判定した場合には、文書検索手段14は、入力語リストの中から複数の入力語を再度選択し、その複数の入力語を全て含む文書を文書群15の中から検索し、検索結果を検索結果文書群18に追加する処理を繰り返す。
- [0134] 文書検索手段14は、例えば、入力語リストの中から複数の入力語を選択した回数が所定に達した場合に十分な検索結果文書群18が得られたと判定し、複数の入力語を選択した回数が所定の回数未満である場合に十分な検索結果文書群18が得られていないと判定すればよい。あるいは、文書検索手段14は、検索結果文書群18に含まれる文書数(レコード数)が所定数に達した場合に十分な検索結果文書群18が得られたと判定し、検索結果文書群18に含まれる文書数が所定数未満である場合に十分な検索結果文書群18が得られていないと判定してもよい。ここでは2種類の判定方法を示したが、文書検索手段14は、後者の判定方法(検索結果文書群18に含まれる文書数に基づく判定方法)で判定を行うことが好ましい。検索結果文書群18に含まれる文書数をより多くすることができ、その結果、文書から抽出される語句の信頼性を向上させることができるからである。なお、十分な検索結果文書群18が得られたか否かの判定に用いるしきい値の情報は、文書検索手段14が予め保持していてもよい。あるいは、入力装置100を介して利用者から文書検索手段14に入力されてもよい。
- [0135] 文書検索手段14によって十分な検索結果文書群18が得られたと判定した場合、データ処理装置200のパターン候補作成手段11は、第1の実施の形態におけるステップS1～S4(図2参照。)と同様の処理を行う。ただし、パターン候補作成手段11

は、1レコード分のデータ(1つの文書IDおよびその文書IDに対応する文書本文)を、文書群15ではなく、検索結果文書群18から取り出して、ステップS1～S4と同様の処理を行う。また、パターン候補作成手段11は、ステップS4において、検索結果文書群18に含まれる全ての文書に対してステップS1～S3の処理を行ったか否かを判定する。ステップS1～S3の処理を実行していない文書が検索結果文書群18に残っている場合、ステップS1に移行してステップS1以降の処理を繰り返す。

[0136] 検索結果文書群18に含まれる全ての文書に対してステップS1～S3の処理を行ったならば、語句候補作成手段12は、第1の実施の形態におけるステップS5～S7と同様の処理を行う。ただし、語句候補作成手段12は、パターン候補16の記録に含まれる文書IDに対応する文書本文を読み込む場合、文書群15ではなく、検索結果文書群18から読み込む。

[0137] パターン候補16の全ての記録に対してステップS5, S6の処理を実行したならば(ステップS7のY)、語句選択手段13は、第1の実施の形態におけるステップS8以降と同様の処理を行う。

[0138] 情報抽出システムの利用者は、出力語リストに含まれる出力語と自身が入力した入力語とを含む辞書を作成すればよい。また、利用者は、各出力語に対応する語句スコアを参照して、個々の出力語を辞書に含めるべきか否かを精査してもよい。

[0139] 次に、第3の実施の形態の効果について説明する。本実施の形態では、同種の語句は同一文書に出現することが多いという性質を利用し、文書検索手段14が、複数の入力語を含む文書を文書群15から検索して、検索結果を検索結果文書群18として記憶させる。そして、パターン候補作成手段11および語句候補作成手段12は、文書群15の代わりに検索結果文書群18を用いて第1の実施の形態と同様の処理を行う。従って、同種の語句が出現する可能性の高い文書のみを対象に処理するため、信頼性の高い語句を出力できる。また、第1の実施の形態と同様の効果も得られる。

[0140] また、上記の第3の実施の形態において、パターン候補作成手段11、語句候補作成手段12、語句選択手段13は、第2の実施の形態と同様の動作を行ってよい(ただし、文書群15ではなく、検索結果文書群18を用いる)。その場合には、第2の実施の形態と同様の効果が得られる。

[0141] 実施の形態4.

図13は、本発明による情報抽出システムの第4の実施の形態を示すブロック図である。第3の実施の形態と同様の構成部については、図12と同一の符号を付し、説明を省略する。第4の実施の形態では、データ処理装置200は、パターン候補作成手段11、語句候補作成手段12、語句選択手段13、文書検索手段14に加え、再作成手段19も含む。

[0142] 再作成手段19は、出力語リストをもとに入力語リストを再作成する。具体的には、再作成手段19は、語句選択手段13から出力語リストを受け取り、出力語リストに含まれる出力語の集合を入力語リストとして文書検索手段14に渡す。このとき、再作成手段19は、記憶装置400に記憶されている検索結果文書群18、パターン候補16、および語句候補17を削除する。すなわち、再作成手段19が受け取った出力語リストが作成されたときに記憶装置400に記憶された検索結果文書群18、パターン候補16、および語句候補17を削除する。

[0143] また、再作成手段19は、語句選択手段13から受け取った出力語リストの複製を内部に蓄積する。

[0144] 再作成手段19が文書検索手段14に入力語リストを渡した後、文書検索手段14、パターン候補作成手段11、語句候補作成手段12、語句選択手段13は、第3の実施の形態と同様の処理を実行する。ただし、語句選択手段13は、出力語リストを再作成手段19に渡す。再作成手段19は、十分な出力語リストが得られたならば、その出力語リストを出力させる。

[0145] 次に、本実施の形態における処理経過の例について説明する。利用者によって、入力装置100に入力語リストが入力される。入力装置100は、入力された入力語リストをデータ処理装置200に送る。すなわち、データ処理装置200には、入力装置100から入力語リストが入力される。

[0146] その後、文書検索手段14、パターン候補作成手段11、語句候補作成手段12、および語句選択手段13は、第3の実施の形態と同様の処理を行ない出力語リストを作成する。語句選択手段13は、作成した出力語リストを再作成手段19に渡す。

[0147] 再作成手段19は、語句選択手段13から出力語リストを受け取ると、十分な出力語リ

ストが蓄積されたか否かを判定する。再作成手段19は、例えば、これまでに蓄積してきた出力語リストの複製に含まれる出力語の総数(ただし、重複する出力語はカウントしない。)が所定数未満であれば十分な出力語リストが蓄積されていないと判定し、これまでに蓄積してきた出力語リストの複製に含まれる出力語の総数が所定数以上であれば十分な出力語リストが蓄積されていると判定してもよい。

[0148] また、例えば、再作成手段19は、前回、語句選択手段13から出力語リストを受け取った時点で蓄積していた出力語リストの複製に含まれる出力語の総数(ただし、重複する出力語はカウントしない。)に対する、新たに受け取った出力語リストに含まれる新たな出力語(それまでに蓄積されていた出力語とは重複しない新たな出力語)の割合(すなわち、出力語の増加率)を計算してもよい。そして、再作成手段19は、その増加率が所定値を越えていれば十分な出力語リストが蓄積されていないと判定し、その増加率が所定値以下であれば十分な出力語リストが蓄積されていると判定してもよい。

[0149] また、例えば、再作成手段19は、出力語リストに含まれる出力語の集合を入力語リストとして文書検索手段14に渡した回数が所定回数以下であれば、十分な出力語リストが蓄積されていないと判定し、その回数が所定回数に達したならば、十分な出力語リストが蓄積されていると判定してもよい。

[0150] ここでは、3種類の判定方法を示したが、この判定方法は例示であり、他の判定方法によって判定を行ってもよい。また、上記の3種類の各判定方法においてしきい値として用いる値の情報は、再作成手段19が予め保持していてもよい。あるいは、入力装置100を介して利用者から再作成手段19に入力されてもよい。

[0151] また、特に、上述の3種類の判定方法のうち、出力語の増加率に基づく判定方法によって判定を行うことが好ましい。新たに出現する出力語の増加率が低いということは、既に検索結果文書群18から網羅的に出力語を収集できていると考えられるからである。また、この判定方法においてしきい値となる所定の値は例えば10%程度とすることが好ましい。例えば、増加率が10%をこえているときには十分な出力語リストが蓄積されていないと判定し、増加率が10%以下のときには十分な出力語リストが蓄積されていると判定することが好ましい。

[0152] 再作成手段19は、十分な出力語リストが蓄積されていないと判定した場合、語句選択手段13から受け取った出力語リストの複製を内部に蓄積する。そして、語句選択手段13から受け取った出力語リストに含まれる出力語の集合を入力語リストとして文書検索手段14に渡す。このとき、再作成手段19は、記憶装置400に記憶されている検索結果文書群18、パターン候補16、および語句候補17を削除する。再作成手段19によって出力語の集合が入力語リストとして文書検索手段14に渡されると、再び、文書検索手段14、パターン候補作成手段11、語句候補作成手段12、および語句選択手段13は、第3の実施の形態と同様の処理を行ない出力語リストを作成する。語句選択手段13は、作成した出力語リストを再作成手段19に渡し、再作成手段19は、上述の動作を繰り返す。

[0153] また、再作成手段19は、十分な出力語リストが蓄積されていると判定した場合、これまで内部に蓄積していた出力語リストを出力装置300に出力させる。このとき、内部に蓄積していた出力語リスト内に重複する語句が存在する場合、語句スコアをまとめることで語句の重複を排除して、語句が重複しないようにして出力語リストを出力装置300に出力させる。語句スコアをまとめて語句の重複を排除する場合、重複する語句の語句スコアの平均値、合計値、最大値、あるいは最小値を求めて、求めた値をその語句の語句スコアとして、重複していた語句およびその語句スコアを出力語リストから除外すればよい。特に、重複する語句の語句スコアの合計値を、その語句の語句スコアとすることが好ましい。複数回出力語となった語句は辞書に含めるべき語句としての信頼度が高いと考えられ、語句スコアの合計値を重複する語句の語句スコアとすることで、そのような信頼度が高いと考えられる語句の語句スコアを高めることができるからである。

[0154] 情報抽出システムの利用者は、出力装置300から出力された出力語リストに含まれる出力語と自身が入力した入力語とを含む辞書を作成すればよい。また、利用者は、各出力語に対応する語句スコアを参照して、個々の出力語を辞書に含めるべきか否かを精査してもよい。

[0155] 次に、第4の実施の形態の効果について説明する。本実施の形態では、新たに見つかった同種の語句をさらに入力語リストとして用いて出力語リストに含める出力語の

数を増加させることができる。従って、同種の語句を網羅的にまとめた辞書を作成することができる。また、上記の第4の実施の形態において、パターン候補作成手段11、語句候補作成手段12、語句選択手段13は、第2の実施の形態と同様の動作を行ってよい(ただし、文書群15ではなく、検索結果文書群18を用いる)。その場合には、第2の実施の形態と同様の効果が得られる。

[0156] また、上記の第4の実施の形態では、再作成手段19が出力語リストに含まれる出力語の集合を入力語リストとして文書検索手段14に渡す場合を示した。第4の実施の形態において、データ処理装置200は、文書検索手段14を備えずに検索結果文書群18を作成しない構成であってもよい。その場合、再作成手段19が出力語リストに含まれる出力語の集合を入力語リストとしてパターン候補作成手段11に渡し、その後のデータ処理装置200の動作では、文書群15に含まれる文書を処理対象として処理を実行すればよい。

[0157] ここで、第1の実施の形態から第4の実施の形態までの各情報抽出システムの具体的な構成例について説明する。図14は、上述の各実施の形態の各情報抽出システムの具体的な構成例を示すブロック図である。図14に示す例において、データ処理装置200は、プログラムに従って動作するコンピュータである。データ処理装置200には、キーボードやマウス等の入力装置100と、ディスプレイ装置あるいはプリンタ等の出力装置300が接続される。また、データ処理装置200には、記憶装置400が接続される。記憶装置400は、文書群15、パターン候補16、語句候補17等を記憶する記憶装置であり、データ処理装置とバスなどで接続されていてもよいし、あるいは、通信ネットワークを介して接続されていてもよい。第3、第4の実施の形態の場合、記憶装置400は、検索結果文書群18も記憶する。また、データ処理装置200は、情報抽出プログラム500を記憶するプログラム記憶装置501を備える。データ処理装置200は、プログラム記憶装置501から情報抽出プログラム500を読み込み、情報抽出プログラム500に従って動作する。この結果、データ処理装置200は、パターン候補作成手段11、語句候補作成手段12、語句選択手段13、および語句選択手段13として動作する。また、第3の実施の形態の場合、文書検索手段14としての動作も行う。第4の実施の形態の場合、再作成手段19としての動作も行う。また、コンピュータで

あるデータ処理装置200は内部に記憶装置を備え、その記憶装置に情報(例えば、出力語リストの複製)を記憶してもよい。

[0158] また、上述の各実施の形態において、各手段(パターン候補作成手段11、語句候補作成手段12、語句選択手段13、文書検索手段14、再作成手段19)をそれぞれ別個のハードウェア装置として、データ処理装置200が備えていてもよい。

[0159] また、上記の各実施の形態において、データ処理装置200は、出力語リストに含まれる出力語と、入力語リストに含まれる入力語とをあわせることによって、辞書を作成してもよい。すなわち、データ処理装置200が自動的に辞書を作成してもよい。

[0160] また、上記の各実施の形態では、入力装置100の例としてキーボードやマウスを挙げたが、通信ネットワークを介して他の装置から情報抽出システムに入力語リストが入力されてもよい。この場合、通信ネットワークとの通信インタフェースを入力装置100として用いればよい。また、出力語リストの出力態様も、通信ネットワークを介して他の装置に出力語リストを出力する態様であってもよい。この場合も、通信ネットワークとの通信インタフェースを出力装置300として用いればよい。

[0161] なお、入力手段は、入力装置100によって実現される。パターン決定手段は、パターン候補作成手段11によって実現される。語句候補抽出手段は、語句候補作成手段12によって実現される。語句選択手段は、語句選択手段13によって実現される。文書検索手段は、文書検索手段14によって実現される。再作成手段は、再作成手段19によって実現される。

[0162] 実施の形態5.

次に、本発明の第5の実施の形態について説明する。図15は、本発明による情報サービスシステムの構成例を示すブロック図である。本実施の形態の情報サービスシステムは、情報抽出システム1000と、辞書サービスシステム2000と、辞書データベース3000とを備える。

[0163] 情報抽出システム1000は、第1の実施の形態から第4の実施の形態のうちのいずれかの情報抽出システムである。ただし、本実施の形態では、入力装置100および出力装置300(図15において図示せず。)は、辞書サービスシステム2000との通信インタフェースによって実現される。情報抽出システム1000には辞書サービスシステ

ム2000から入力語リストが入力され、入力語と同種の出力語のリストを辞書サービスシステム2000に出力する。

[0164] 辞書データベース3000は、各種辞書3001, 3002を記憶する記憶装置である。各種辞書3001, 3002は、情報抽出システム1000が出力した出力語およびその出力語の抽出のために入力語リスト作成者によって作成された入力語を含む。また、各辞書は、その辞書の種類の情報と対応付けて登録される。

[0165] 辞書サービスシステム2000には入力語リスト作成者から入力語リストが入力され、辞書サービスシステム2000はその入力語リストを用いて情報抽出システム1000に出力語リストを出力させる。そして、出力語リストを入力語リスト作成者に提示して、入力語リスト作成者に入力語および出力語を含む辞書の作成を促す。辞書サービスシステム2000は、入力語リスト作成者から辞書を受け取ると、その辞書の種類の情報と対応付けて、辞書および辞書の種類の情報を辞書データベース3000に登録する。また、辞書サービスシステム2000は、辞書購入者の要求に応じて、辞書データベース3000に登録されている辞書を辞書購入者に提供する。また、辞書サービスシステム2000は、辞書サービス運営者が辞書購入者から受け取るべき金額や、入力語リスト作成者に支払うべき金額を記録する。

[0166] 以下の説明では、辞書サービスシステム2000が辞書購入者の端末(以下、購入者端末と記す。)および入力語リスト作成者の端末(以下、作成者端末と記す。)と情報を送受信するものとして説明する。購入者端末は、辞書購入者によって操作される端末であり、作成者端末は、入力語リスト作成者によって操作される端末である。

[0167] 図16は、辞書サービスシステム2000の構成例を示すブロック図である。辞書サービスシステム2000は、制御部62と金銭情報記憶手段61とを備える。制御部62は、辞書サービスシステム2000が備える記憶装置(図示せず。)に記憶されたプログラムに従って動作する。制御部62は、購入者端末51、作成者端末52、情報抽出システム1000、辞書データベース3000との間で情報を授受する。なお、辞書サービスシステム2000は、購入者端末51、作成者端末52、情報抽出システム1000との通信を行う際に情報を送受信する通信インタフェースを備えるが、図16では、通信インタフェースの図示を省略している。また、辞書サービスシステム2000は、辞書データベー

ス3000への情報の書き込みや読み込みを行うためのインタフェースを備えるが、図16では、そのインタフェースの図示を省略している。制御部62は、通信インタフェース(図示せず。)を介して他の装置と情報を送受信したり、辞書データベース3000とのインタフェース(図示せず。)を介して情報の読み書きを行う。

[0168] また、金銭情報記憶手段61は、辞書サービス運営者が入力語リスト作成者に支払う金額や、辞書購入者から受け取る金額を記憶する。制御部62は、これらの金額を金銭情報記憶手段61に記憶させる。辞書サービス運営者は、辞書サービスシステム2000、情報抽出システム1000、および辞書データベース3000の管理者である。

[0169] 次に、動作について説明する。

購入者端末51は、辞書購入者の操作に従って、辞書購入者が購入を希望する辞書の種類を辞書サービスシステム2000に送信する。辞書サービスシステム2000の制御部62は、その種類の情報を受信する。

[0170] 続いて制御部62は、辞書データベース3000を検索して、辞書購入者の希望する種類の辞書が辞書データベース3000に登録されているか否かを判定する。

[0171] 辞書購入者の希望する種類の辞書が辞書データベース3000に登録されていないと判定した場合、制御部62は、以下のように動作する。制御部62は、辞書購入者が購入を希望する辞書の種類を作成者端末52に送信することによって、辞書購入者の希望する辞書の種類を入力語リスト作成者に提示する。

[0172] 入力語リスト作成者は、その種類に応じた入力語リストを作成し、辞書サービスシステム2000に渡す。このとき、作成者端末52には、入力語リスト作成者によって作成された入力語リストが入力され、作成者端末52は、入力語リスト作成者の操作に従って、その入力語リストを辞書サービスシステム2000に送信する。辞書サービスシステム2000の制御部62は、その入力語リストを受信する。

[0173] 制御部62は、入力語リストを受信すると、その入力語リストを作成した入力語リスト作成者に対して辞書サービス運営者が支払うべき金額(辞書作成の対価)の情報を、金銭情報記憶手段61に記憶させる。このとき、制御部62は、金額の情報と、入力語リスト作成者の識別情報とを対応付けて金銭情報記憶手段61に記憶させる。なお、例えば、作成者端末52が入力語リストを送信するときに、入力語リスト作成者の操

作に従って入力語リスト作成者の識別情報も制御部62に送信することによって、制御部62は入力語リスト作成者の識別情報を得ることができる。

[0174] さらに、制御部62は、作成者端末52から受信した入力語リストを、情報抽出システム1000に出力する。情報抽出システム1000は、制御部62から入力された入力語リストを用いて、出力語リストを作成する。情報抽出システム1000は、第1から第4のいずれかの実施の形態で説明した動作によって出力語リストを作成すればよい。情報抽出システム1000は、辞書サービスシステム2000に出力語リストを出力し、制御部62はこの出力語リストを得る。

[0175] 制御部62は、出力語リストを作成者端末52に送信して、入力語リスト作成者に辞書の作成を促す。入力語リスト作成者は、自身が作成した入力語リストに含まれる入力語と制御部62から受信した出力語リストに含まれる出力語を含む辞書を作成する。このとき、入力語リスト作成者は、出力語リストを精査して辞書を作成してもよい（例えば、不要な出力語を除外して辞書を作成してもよい。）。作成者端末52は、作成された辞書を入力され、入力語リスト作成者の操作に従って、その辞書を辞書サービスシステム2000に送信する。制御部62は、辞書を受信すると、辞書購入者が希望した辞書の種類と辞書とを対応付けてデータベース3000に登録する。

[0176] 次に、制御部62は、辞書購入者が希望した種類の辞書を辞書データベース3000から読み込んで、購入者端末51に送信することによって、辞書を辞書購入者に提供する。

[0177] 辞書購入者の希望する種類の辞書が辞書データベース3000に登録されていると判定した場合、制御部62は、その辞書を辞書データベース3000から読み込み、購入者端末51に送信することによって、辞書を辞書購入者に提供すればよい。

[0178] 制御部62は、辞書を辞書データベース3000から読み込んで購入者端末51に送信した場合、辞書購入者の識別情報と、辞書サービス運営者が辞書購入者から辞書の対価として受け取る金額の情報とを金銭情報記憶手段61に記憶させる。なお、例えば、購入者端末51が辞書の種類を送信するときに、辞書購入者の操作に従って辞書購入者の識別情報も制御部62に送信することによって、制御部62は辞書購入者の識別情報を得ることができる。

[0179] 辞書サービス運営者は、金銭情報記憶手段61に記憶された辞書購入者の識別情報および金額を確認し、辞書購入者に辞書の対価を請求する。辞書購入者は、その対価を辞書サービス運営者に支払う。また、辞書サービス運営者は、金銭情報記憶手段61に記憶された入力語リスト作成者の識別情報および金額を確認し、その金額を入力語リスト作成者に支払う。

[0180] ここで、入力語リスト作成者と辞書購入者が同じであってもよい。その場合、入力語リストの作成者に対して対価を支払う必要はない。

[0181] 第5の実施の形態によれば、これまでよりも安価に辞書を提供できる。その理由は、情報抽出システムによって自動的に辞書を作成できるためである。従来、辞書を販売するためには、大量の文書から人手などにより辞書を作成していたためコストがかかっていた。情報抽出システムを用いることで、辞書を自動的に作成できるため、これまでより安価に提供できる。

[0182] また、辞書サービス運営者が入力語リストを作成してもよいが、その場合、辞書サービス運営者に、要求された辞書の種類に関する知識が必要である。本実施の形態において、入力語リスト作成者を広く募集し、入力語リスト作成者に対価を支払うことで、幅広い種類の辞書作成に対応できる。

[0183] 実施の形態6.

次に、本発明の第6の実施の形態について説明する。本実施の形態の情報サービスシステムの構成は、第5の実施の形態と同様である(図15および図16参照。)。ただし、入力語リスト作成者に辞書作成の対価を支払う態様が異なる。本実施の形態では、辞書を辞書購入者に販売(提供)した場合に、辞書作成の対価として辞書サービス運営者が支払うべき金額を金銭情報記憶手段61に記憶する。すなわち、辞書が販売されることによって、入力語リスト作成者に辞書作成の対価が支払われることになる。また、入力語リスト作成者に作成料を支払うとき、辞書サービスシステム利用料としての対価を差し引く。

[0184] 以下の説明においても、辞書サービスシステム2000が辞書購入者の端末(購入者端末)および入力語リスト作成者の端末(作成者端末)と情報を送受信するものとして説明する。

[0185] 本実施の形態の動作について説明する。

まず、入力語作成者は、入力語リストを作成し、作成者端末52を用いて、入力語リストを辞書サービスシステム2000に送信する。このとき、作成者端末52には、入力語リストが入力され、作成者端末52は、入力語作成者の操作に従って、入力語リストを辞書サービスシステム2000に送信する。辞書サービスシステム2000の制御部62は、入力語リストを受信する。

[0186] 制御部62は、作成者端末52から受信した入力語リストを、情報抽出システム1000に出力する。情報抽出システム1000は、制御部62から入力された入力語リストを用いて、出力語リストを作成する。情報抽出システム1000は、第1から第4のいずれかの実施の形態で説明した動作によって出力語リストを作成すればよい。情報抽出システム1000は、辞書サービスシステム2000に出力語リストを出力し、制御部62はこの出力語リストを得る。

[0187] 制御部62は、出力語リストを作成者端末52に送信して、入力語リスト作成者に辞書の作成を促す。入力語リスト作成者は、自身が作成した入力語リストに含まれる入力語と制御部62から受信した出力語リストに含まれる出力語を含む辞書を作成する。このとき、入力語リスト作成者は、出力語リストを精査して辞書を作成してもよい(例えば、不要な出力語を除外して辞書を作成してもよい。)。入力語リスト作成者は、作成した辞書およびその辞書の種類を作成者端末52から辞書サービスシステム2000に送信させる。すなわち、作成者端末52は、作成された辞書およびその種類を入力され、入力語リスト作成者の操作に従って、辞書およびその種類を辞書サービスシステム2000に送信する。

[0188] 辞書サービスシステム2000の制御部62は、作成者端末52から辞書およびその種類を受信すると、辞書と辞書の種類とを対応付けてデータベース3000に登録する。

[0189] 次に、辞書購入者が、辞書サービスシステム2000を介し、辞書データベース3000を参照し、購入したい種類の辞書を購入する。このとき、購入者端末51は、辞書購入者の操作に従って、制御部62に、辞書購入を希望する旨の情報を送信する。制御部62は、この情報を受信すると、辞書データベース3000に登録されている各辞書の種類を読み込んで、各辞書の種類の情報を購入者端末51に送信し、辞書購入者

に種類の選択を促す。購入者端末51は、辞書購入者の操作に従って、辞書購入者が希望する種類を制御部62に送信する。制御部62は、辞書の種類の情報を購入者端末51から受信すると、その種類に対応する辞書を辞書データベース3000から読み込み、その辞書を購入者端末51に送信する。この結果、辞書購入者に辞書が提供される。また、制御部62は、辞書の対価となる金額と、辞書購入者の識別情報とを対応させて、金銭情報記憶手段61に記憶させる。なお、例えば、購入者端末51が辞書購入者の識別情報を送信することによって、制御部62に辞書購入者の識別情報を通知すればよい。

[0190] また、制御部62は、辞書を辞書購入者に提供したときに（辞書を購入者端末51に送信したときに）、辞書作成の対価として辞書サービス運営者が入力語リスト作成者に支払うべき金額と、その入力語リスト作成者の識別情報とを対応させて、金銭情報記憶手段61に記憶させる。なお、例えば、作成者端末52が、辞書を送信するときに入力語リスト作成者の識別情報もあわせて送信することによって、制御部62に入力語リスト作成者の識別情報を通知すればよい。

[0191] 辞書サービス運営者は、金銭情報記憶手段61に記憶された辞書購入者の識別情報および金額を確認し、辞書購入者に辞書の対価を請求する。辞書購入者は、その対価を辞書サービス運営者に支払う。

[0192] また、辞書サービス運営者は、金銭情報記憶手段61に記憶された入力語リスト作成者の識別情報および金額を確認し、その金額を入力語リスト作成者に支払う。このとき、辞書サービス運営者は、辞書サービスシステム利用料を差し引いた額を入力語リスト作成者に支払う。なお、制御部62は、辞書作成の対価から辞書サービスシステム利用料を差し引いた額を金銭情報記憶手段61に記憶させておいてもよい。

[0193] また、入力語リスト作成者と辞書購入者が同じであってもよい。その場合、入力語リスト作成者に対して辞書作成の対価を支払う必要はなく、辞書購入者は辞書サービス運営者に対し辞書サービスシステム利用料を支払うのみでよい。この場合、制御部62は、辞書提供の対価の代わりに、辞書サービスシステム利用料を金銭情報記憶手段61に記憶させればよい。

[0194] 本実施の形態では、コストを抑えながらも辞書の信頼性を簡単に保つことができる。

従来の辞書作成では、文書から人手で辞書項目となる語句を収集していたため、信頼性は高いがコストがかかっていた。情報抽出システムを用いることで、出力語リストに辞書項目の候補となる語句を自動的に出力できるため、辞書作成のコストを抑えることができる。また、出力語リストを精査することで辞書の信頼性を保つことができる。

[0195] なお、登録手段、辞書提供手段、金額情報登録手段は、辞書サービスシステム2000の制御部62によって実現される。金銭情報記憶手段は、辞書サービスシステム2000の金銭情報記憶手段61によって実現される。

[0196] 実施の形態7.

次に、本発明の第7の実施の形態について説明する。図17は、本実施の形態の情報サービスシステムの構成例を示すブロック図である。第6の実施の形態と同様の構成部については、図15と同一の符号を付し、説明を省略する。本実施の形態の情報サービスシステムは、トレンドキーワードデータベース4000を備える。なお、辞書サービスシステム2000は、第6の実施の形態と同様に、制御部62および金銭情報記憶手段61(図16参照。)を備える。

[0197]トレンドキーワードデータベース4000は、流行となっているキーワードのリストを記憶する記憶装置である。本実施の形態では、辞書サービスシステム2000の制御部62(図17において図示せず。図16参照。)は、辞書データベース3000に対する読み書きと同様に、トレンドキーワードデータベース4000に対する情報の読み書きも行う。

[0198] 制御部62は、流行となっているキーワードのリストを作成し、トレンドキーワードデータベース4000に記憶させる。制御部62は、インターネットで一般的に利用されている検索エンジンの検索キーワードログにアクセスして、数多く検索されている語をキーワードとして抽出し、そのキーワードのリストをトレンドキーワードデータベース4000に記憶させればよい。検索エンジンで検索された語を検索回数の多い順にランク付けして公表しているWebページがある。制御部62は、このようなWebページの表示データを取得して、そのWebページで検索回数の上位にランク付けされているキーワードを抽出すればよい。このようなWebページのURLとして、例えば、「<http://guide.search.goo.ne.jp/ranking/>」、「<http://picks.dir.yahoo.co.jp/new/review2005/index.h>

tml」、「<http://search.biglobe.ne.jp/ranking.html>」等がある。また、制御部62は、ニュース記事を公表しているWebページの表示データを取得して、そのWebページに登場する語をキーワードとして取得してもよい。また、キーワードのリストを人手で作成し、辞書サービスシステム2000に入力して、トレンドキーワードデータベースに記憶させてもよい。

[0199] 次に、動作について説明する。

まず、入力語リスト作成者が、辞書サービスシステム2000を介し、トレンドキーワードデータベース4000を参照する。例えば、作成者端末52(図17において図示せず。図16参照。)が、入力語リスト作成者の操作に従って、トレンドキーワードデータベース4000に記憶されているキーワードのリストを、辞書サービスシステム2000に要求する。辞書サービスシステム2000の制御部62は、この要求に応じて、トレンドキーワードデータベース4000に記憶されているキーワードのリストを読み込み、作成者端末52に送信する。作成者端末は、受信したキーワードのリストを表示して、入力語リスト作成者にキーワードのリストを提示する。

[0200] キーワードのリストは、流行となっている語句のリストであるので、キーワードのリストによって、販売の可能性の高い辞書の種類を判断することができる。入力語リスト作成者は、キーワードのリストを参照して、そのような辞書の種類を判断し、その辞書に含めるべき入力語のリストを作成する。このように、入力語リスト作成者は、販売の可能性の高い種類の辞書に応じた入力語リストを作成する。

[0201] 以降の動作は、第6の実施の形態の動作と同様である。この動作によって、販売の可能性の高い辞書(すなわち、売れる可能性の高い辞書)を辞書データベース3000に登録することができる。

[0202] 本実施の形態では、入力語リスト作成者が、トレンドキーワードデータベース4000に記憶されているキーワードを参照することによって、どのような種類の語句を含む辞書が売れそうか判断することができる。従って、入力語リスト作成者は、販売の可能性の高い辞書の種類を容易に判断することができる。また、入力語リスト作成者がそのような辞書を作成するための入力語リストを作成すれば、その入力語リストを用いて、情報抽出システムが迅速に出力語リストを作成する。従って、入力語リスト作成者は、

販売の可能性の高い辞書をすばやく作成することができる。また、作成された辞書は販売の可能性が高いと考えられるので、辞書の販売数を増加し、入力語リスト作成者と、辞書サービス運営者の収入が増加する。

[0203] なお、流行キーワード記憶手段は、トレンドキーワードデータベース4000によって実現される。流行キーワード提供手段は、辞書サービスシステム2000の制御部62によって実現される。

[0204] 実施の形態8.

次に、本発明の第8の実施の形態について説明する。図18は、第8の実施の形態の情報サービスシステムの構成例を示すブロック図である。本実施の形態の情報サービスシステムは、情報抽出システム1000と、広告サービスシステム5000と、辞書データベース3000とを備える。また、広告サービスシステム5000には、通信ネットワークを介して広告閲覧システム6000が接続される。

[0205] 情報抽出システム1000は、第1の実施の形態から第4の実施の形態のうちのいずれかの情報抽出システムである。ただし、本実施の形態では、入力装置100および出力装置300(図18において図示せず。)は、広告サービスシステム5000との通信インタフェースによって実現される。情報抽出システム1000には広告サービスシステム5000から入力語リストが入力され、入力語と同種の出力語のリストを広告サービスシステム5000に出力する。情報抽出システム1000は、情報抽出システム運営者によって運営される。

[0206] 広告サービスシステム5000は、広告主から広告とその広告に関連する関連キーワードを受け取り、広告と関連キーワードとを対応付けて広告サービスシステム5000内に記憶する。また、広告閲覧システム6000からキーワードを受け取り、このキーワードに関連する広告を返す。このとき、辞書データベース3000を参照し、キーワードを含んでいる辞書がある場合、その辞書内のキーワード(語句)を取得し、取得したキーワードに対応する広告も広告閲覧システムに返す。このようにして広告サービスシステム5000は、広告閲覧システム6000に広告を配信する。広告サービスシステム5000は、広告サービス運営者によって運営される。

[0207] 辞書データベース3000は、各種辞書3001、3002を記憶する記憶装置である。各

種辞書3001, 3002は、情報抽出システム1000が出力した出力語およびその出力語の抽出のために広告サービス運営者によって作成された入力語を含む。また、各辞書は、その辞書の種類の情報と対応付けて記憶される。

[0208] 広告閲覧システム6000は、広告閲覧者からキーワードや文書を受け取る。広告閲覧システム6000は、文書を受け取った場合は、一般的な形態素解析ツールなどを用いて文書内の文字列を単語に分割し、分割によって得られた単語をキーワードとする。そして、そのキーワードを広告サービスシステム5000に渡し、キーワードに対応した広告を受け取り、広告を表示する。表示には一般的な文書閲覧ツールやWebブラウザなどを用いる。

[0209] 以下の説明では、広告サービスシステム5000が広告主の端末(以下、広告主端末)および広告サービス運営者の端末(以下、運営者端末)と情報を送受信するものとして説明する。広告主端末は、広告主によって操作される端末であり、運営者端末は、広告サービス運営者によって操作される端末である。

[0210] 図19は、広告サービスシステム5000と広告閲覧システム6000の構成例を示すブロック図である。広告サービスシステム5000は、制御部72と広告記憶手段71とを備える。制御部72は、広告サービスシステム5000が備える記憶装置(図示せず。)に記憶されたプログラムに従って動作する。制御部72は、広告主端末55、運営者端末56、情報抽出システム1000、辞書データベース3000との間で情報を授受する。なお、広告サービスシステム5000は、広告主端末55、運営者端末56、情報抽出システム1000との通信を行う際に情報を送受信する通信インタフェースを備えるが、図19では、通信インタフェースの図示を省略している。また、広告サービスシステム5000は、辞書データベース3000への情報の書き込みや読み込みを行うためのインタフェースを備えるが、図19ではそのインタフェースの図示を省略している。制御部72は、通信インタフェース(図示せず。)を介して他の装置と情報を送受信したり、辞書データベース3000とのインタフェース(図示せず。)を介して情報の読み書きを行う。

[0211] また、広告記憶手段71は、広告と、その広告に関連する関連キーワードとを対応付けて記憶する。制御部72は、広告主端末55から広告および関連キーワードを受信し、広告と関連キーワードとを対応付けて広告記憶手段71に記憶させる。

- [0212] 広告閲覧システム6000は、文書等入力手段81と、広告要求手段82と、広告表示手段83とを備える。文書等入力手段81は、キーワードや文書が入力されるキーボード等の入力装置である。広告表示手段83は、広告を表示するディスプレイ装置である。
- [0213] 広告要求手段82は、プログラム(形態素解析プログラムやブラウザ等の各種プログラム)に従って動作する。広告要求手段82は、文書等入力手段81に入力されたキーワードを広告サービスシステム5000に送信し、そのキーワードに関連する広告を要求する。また、広告要求手段82は、文書等入力手段81に文書が入力された場合、その文書内の文字列に対して形態素解析を行って単語に分割し、その単語をキーワードとする。広告要求手段82は、広告サービスシステム5000から広告を受信すると、その広告を広告表示手段83に表示させる。
- [0214] 次に、動作について説明する。本実施の形態の動作は、大きく次の3つのフェーズに分かれる。第1のフェーズは、広告サービスシステム5000が広告主(広告主端末55)から広告を受け取り、広告記憶手段71に登録するフェーズである。第2のフェーズは、広告サービスシステム5000が情報抽出システム1000に出力語リストを作成させて、辞書を辞書データベースに登録するフェーズである。第3のフェーズは、広告閲覧システム6000が広告サービスシステム5000から広告を受信して広告を表示するフェーズである。第1のフェーズおよび第2のフェーズは非同期で行われ、その後、第3のフェーズが行われる。
- [0215] 第1のフェーズについて説明する。広告主端末55は、広告主から入力された広告および関連キーワードを、広告主の操作に従って広告サービスシステム5000に送信する。広告サービスシステム5000の制御部72は、広告主端末55から広告および関連キーワードを受信する。
- [0216] ここで、広告とは広告主が購入して欲しい商品やサービスの宣伝素材であり、自然文文字列や画像等を含んでいる。関連キーワードとは、広告主が購入して欲しい商品やサービスに関連する語句であり、商品名やサービス名、その商品の特徴をあらわす語句等が関連キーワードに該当する。特に、本実施の形態では、関連キーワードが商品名やサービス名を含んでいることが好ましい。商品名やサービス名はその商

品固有の語句であり最も特徴を現していると考えられるからである。例えば、広告が、セキュリティに関わる商品の広告である場合、「セキュリティ製品X」、「情報漏洩」、「セキュリティ」等を関連キーワードとすればよい。

[0217] 制御部72は、広告と関連キーワードを受信すると、その広告と関連キーワードリストとを対応付けて広告記憶手段71に記憶させる。以上が第1のフェーズの動作である。

[0218] 第2のフェーズについて説明する。まず、広告サービス運営者が、入力語リストを作成する。入力語リストを作成する際、広告サービス運営者は、広告サービスシステム5000の広告記憶手段71に記憶されている広告の関連キーワードリストを参考にしてもよい。そして、関連キーワードリストを参考に、種類の多い商品名やサービス名などから入力語リストを作成してもよい。例えば、セキュリティ関係の製品に関する関連キーワードが多い場合、その製品名「セキュリティ製品X」、「セキュリティ製品Y」、「セキュリティ製品Z」等を入力語とする入力語リストを作成してもよい。

[0219] 運営者端末56は、作成された入力語リストおよびその種類を入力され、広告サービス運営者の操作に従って入力語リストおよび種類の情報を広告サービスシステム5000に送信する。広告サービスシステム5000の制御部72は、入力語リストおよび種類の情報を運営者端末56から受信すると、その入力語リストおよび種類の情報を情報抽出システム1000に送信する。例えば、制御部72は、「セキュリティ製品X」、「セキュリティ製品Y」、「セキュリティ製品Z」という入力語からなる入力語リストと、その種類(本例では「セキュリティ」)を運営者端末56から受信し、その入力語リストおよび種類を情報抽出システム1000に渡す。

[0220] 情報抽出システム1000は、制御部72から入力語リストを受け取ると、その入力語リストを用いて出力語リストを作成する。情報抽出システム1000は、第1から第4のいずれかの実施の形態で説明した動作によって出力語リストを作成すればよい。この結果、例えば、セキュリティ関連製品名のリストが出力語リストとして作成される。情報抽出システム1000は、広告サービスシステム5000に出力語リストを出力し、制御部72はこの出力語リストを得る。

[0221] 広告サービスシステム5000の制御部72は、出力語リストに含まれる出力語と、入

力語リストに含まれる入力語とをあわせて辞書を作成し、その種類名とともに、辞書データベース3000に辞書を記憶させる。以上が第2のフェーズの動作である。

[0222] 第3のフェーズについて説明する。まず、広告閲覧システム6000の文書等入力手段81に広告閲覧者からキーワードリストや文書が入力される。キーワードは、インターネットなどで一般的に使われている検索エンジンに入力されたキーワード等のように、広告閲覧者が興味を持っている語句である。文書等入力手段81に入力される文書は、自然文テキストやHTML文書等のある程度形式が決まった文書などである。

[0223] 広告要求手段82は、文書が入力された場合にのみ、その文書を一般的な形態素解析ツールを用いて単語に分割し、その単語をキーワードリストとする動作を行う。広告要求手段82は、このように作成したキーワードリスト、あるいは広告閲覧者から直接入力されたキーワードを広告サービスシステム5000に送信する。例えば、「セキュリティ製品Y」等のキーワードを広告サービスシステム5000に送信する。

[0224] 広告サービスシステム5000の制御部72は、辞書データベース3000を参照し、広告要求手段82から受信したキーワードリストの語句が、各辞書に含まれているか否かを判定する。キーワードリストの語句が辞書に含まれている場合には、その辞書に含まれている語句群(すなわちキーワードリストの語句と同じ種類の語句群)を辞書データベース3000から読み込む。例えば、「セキュリティ製品Y」という語句が、「セキュリティ」という種類の辞書に含まれている場合、その「セキュリティ」の辞書に含まれている「セキュリティ製品X」、「セキュリティ製品Z」等の語句を得る。

[0225] 次に、制御部72は、広告記憶手段71に記憶されている広告と関連キーワードリストを参照し、広告閲覧システム6000から受け取ったキーワードリスト、および、辞書データベース3000から得られた各語句に対応する広告を、広告記憶手段71から読み込む。例えば、第1のフェーズで記憶した「セキュリティ製品X」の広告等を読み込む。

[0226] 次に、制御部72は、広告記憶手段71から読み込んだ広告を広告閲覧システム6000に送信する。広告閲覧システム6000の広告要求手段82は、この広告を受信すると、広告表示手段82に表示させる。以上が第3のフェーズの動作である。

[0227] 本実施の形態によれば、これまでよりも幅広く、関連する広告を表示でき、その結果、広告閲覧者にとっては選択の幅が広がり、広告主にとっては広告を表示できる回数

が増えるという効果が得られる。その理由は、情報抽出システム1000を用いて、簡単に関連製品などの辞書を作ることができ、この辞書を参照して閲覧する広告を幅広く集めることができるためである。

- [0228] すなわち、広告閲覧システム6000から受け取ったキーワードに関連する広告だけでなく、そのキーワードと同種の語句群を辞書から読み込み、広告閲覧システム6000から受け取ったキーワードに関連する広告および辞書から読み込んだ語句に関連する広告を広告記憶手段71から取得して、広告閲覧システムに提供することができる。例えば、「セキュリティ製品Y」というキーワードを広告閲覧システムから受信した場合、単に「セキュリティ製品Y」の広告を提供するだけでなく、「セキュリティ製品Y」と「セキュリティ製品X」とを同じ種類の語句として含む辞書が辞書データベース3000に登録されていれば、関連製品「セキュリティ製品X」に関する広告も提供することができる。この結果、広告主の広告機会を増やすことができる。

また、製品の移り変わりが頻繁に起こったとしても、情報抽出システム1000が出力語リストを自動的に作成するので、辞書を素早く簡単に作成することができる。

- [0229] なお、登録手段および広告提供手段は、広告サービスシステム5000の制御部72によって実現される。広告記憶手段は、広告サービスシステム5000の広告記憶手段71によって実現される。

産業上の利用可能性

- [0230] 本発明によれば、人名の語句リストや地名の語句リストなどを簡単に作成するといった辞書作成に適用することができる。また、文書中から特定の種類の語句を検索して抽出する情報抽出にも適用することができる。

請求の範囲

- [1] 文書内の語句を文書から抽出して出力する情報抽出システムであって、
複数の語句を含む入力語リストが入力される入力手段と、
形式が異なる文書を含む複数の文書の中から一つの文書を選択し、選択した文書内で、入力語リストに含まれる語句を抽出対象でない文字列と区切る文字列をパターンとして決定することを文書毎に行うパターン決定手段と、
パターンによって区切られる文字列を、当該パターンの決定に用いた文書から抽出して語句の候補とする語句候補抽出手段と、
語句候補抽出手段によって抽出された語句の候補または当該語句の候補に含まれる部分文字列のうち所定の条件を満たす語句の候補または部分文字列を出力対象の語句として選択する語句選択手段とを備えた
ことを特徴とする情報抽出システム。
- [2] パターン決定手段は、入力語リストに含まれる語句の前方の文字列および後方の文字列をパターンとして決定し、
語句候補抽出手段は、前記前方の文字列および前記後方の文字列によって挟まれる文字列を文書から抽出して語句の候補とし、
語句選択手段は、語句候補抽出手段によって抽出された語句の候補の中から出力対象の語句を選択する
請求項1に記載の情報抽出システム。
- [3] 語句選択手段は、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを語句の候補毎に算出し、語句スコアが所定のしきい値以上である語句の候補を出力対象の語句として選択する
請求項2に記載の情報抽出システム。
- [4] 語句選択手段は、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを語句の候補毎に算出し、語句スコアの高い順に語句の候補をソートし所定の順位以上である語句の候補を出力対象の語句として選択する

請求項2に記載の情報抽出システム。

- [5] 語句選択手段は、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを語句の候補毎に算出し、語句スコアの高い順に語句の候補をソートし上位の所定の割合の語句の候補を出力対象の語句として選択する

請求項2に記載の情報抽出システム。

- [6] パターン決定手段は、入力語リストに含まれる語句の前方の文字列のみ、あるいは、入力語リストに含まれる語句の後方の文字列のみをパターンとして決定し、

語句候補抽出手段は、前記前方の文字列に続く文字列、あるいは、前記後方の文字列の直前の文字列を文書から抽出して語句の候補とし、

語句選択手段は、語句の候補に含まれる部分文字列を作成し、作成した各部分文字列の中から出力対象の語句を選択する

請求項1に記載の情報抽出システム。

- [7] 語句選択手段は、語句の候補の総数に対する部分文字列を含む語句の候補の割合を、各語句の候補毎および各部分文字列毎に算出して、前記割合が所定の範囲に属する部分文字列を特定し、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを、特定した部分文字列を含む語句の候補毎に算出して、算出結果を前記部分文字列の語句スコアとし、語句スコアが所定のしきい値以上である部分文字列を出力対象の語句として選択する

請求項6に記載の情報抽出システム。

- [8] 語句選択手段は、語句の候補の総数に対する部分文字列を含む語句の候補の割合を、各語句の候補毎および各部分文字列毎に算出して、前記割合が所定の範囲に属する部分文字列を特定し、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを、特定した部分文字列を含む語句の候補毎に算出して、算出結果を前記部分文字列の語句スコアとし、語句スコアの高い順に部分文字列をソートし所定の順位以上である部分文字列を出力対象の語句として選択する

請求項6に記載の情報抽出システム。

- [9] 語句選択手段は、語句の候補の総数に対する部分文字列を含む語句の候補の割合を、各語句の候補毎および各部分文字列毎に算出して、前記割合が所定の範囲に属する部分文字列を特定し、語句の候補の特定に用いたパターン of 候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを、特定した部分文字列を含む語句の候補毎に算出して、算出結果を前記部分文字列の語句スコアとし、語句スコアの高い順に部分文字列をソートし上位の所定の割合の部分文字列を出力対象の語句として選択する

請求項6に記載の情報抽出システム。

- [10] 形式が異なる文書を含む複数の文書であって、予め記憶装置に記憶されている複数の文書の中から、入力語リストから選択した複数の語句全てを含む文書を検索する文書検索手段を備え、

パターン決定手段は、文書検索手段によって検索された文書毎にパターンの決定を行う

請求項1から請求項9のうちのいずれか1項に記載の情報抽出システム。

- [11] 語句選択手段によって選択された語句のリストを入力語リストとして文書検索手段に付与する再作成手段を備えた

請求項10に記載の情報抽出システム。

- [12] 語句選択手段によって選択された語句のリストを入力語リストとしてパターン決定手段に付与する再作成手段を備えた

請求項1から請求項9のうちのいずれか1項に記載の情報抽出システム。

- [13] 請求項1から請求項12のうちのいずれか1項に記載の情報抽出システムと、

同種の語句のリストである辞書を記憶する辞書データベースと、

辞書の作成および辞書の販売に伴い授受される金額を管理する辞書サービスシステムとを備え、

辞書サービスシステムは、

入力語リスト作成者によって作成された入力語リストを情報抽出システムに入力して、情報抽出システムから語句のリストを受け取り、当該語句のリストおよび前記入力語

リストに含まれる語句を含む辞書を辞書データベースに登録する登録手段と、

辞書購入者の端末からの要求に応じて、辞書データベースに記憶されている辞書を前記端末に送信する辞書提供手段と、

辞書作成および辞書の販売に伴い授受される金額を記憶する金銭情報記憶手段と、

辞書サービスシステムの運用者が入力語リスト作成者に対し辞書の作成の対価として支払うべき金額、および辞書購入者が辞書サービスシステムの運用者に辞書の対価として支払うべき金額を金銭情報記憶手段に記憶させる金額情報登録手段とを含む

ことを特徴とする情報サービスシステム。

[14] 流行している語句を記憶する流行キーワード記憶手段を備え、

辞書サービスシステムは、

流行キーワード記憶手段から流行している語句を読み込んで、入力語リスト作成者の端末に送信する流行キーワード提供手段を含む

請求項13に記載の情報サービスシステム。

[15] 請求項1から請求項12のうちのいずれか1項に記載の情報抽出システムと、

同種の語句のリストである辞書を記憶する辞書データベースと、

広告を表示する広告表示装置からキーワードを受信して、広告表示装置に広告を送信する広告サービスシステムとを備え、

広告サービスシステムは、

入力語リスト作成者によって作成された入力語リストを情報抽出システムに入力して、情報抽出システムから語句のリストを受け取り、当該語句のリストおよび前記入力語リストに含まれる語句を含む辞書を辞書データベースに登録する登録手段と、

広告および当該広告に関連する関連キーワードを対応付けて記憶する広告記憶手段と、

広告表示装置から受信したキーワードを含む辞書を辞書データベースに記憶された辞書の中から検索し、検索された辞書に含まれる語句に対応付けられた広告および広告表示装置から受信したキーワードに対応づけられた広告を広告記憶手段から

読み込んで、広告表示装置に送信する広告提供手段とを含む
ことを特徴とする情報サービスシステム。

- [16] 文書内の語句を文書から抽出する情報抽出方法であって、
入力手段が、複数の語句を含む入力語リストを入力し、
パターン決定手段が、形式が異なる文書を含む複数の文書の中から一つの文書を選択し、選択した文書内で、入力語リストに含まれる語句を抽出対象でない文字列と区切る文字列をパターンとして決定することを文書毎に行い、
語句候補抽出手段が、パターンによって区切られる文字列を、当該パターンの決定に用いた文書から抽出して語句の候補とし、
語句選択手段が、語句候補抽出手段によって抽出された語句の候補または当該語句の候補に含まれる部分文字列のうち所定の条件を満たす語句の候補または部分文字列を出力対象の語句として選択することを特徴とする情報抽出方法。

- [17] パターン決定手段が、入力語リストに含まれる語句の前方の文字列および後方の文字列をパターンとして決定し、
語句候補抽出手段が、前記前方の文字列および前記後方の文字列によって挟まれる文字列を文書から抽出して語句の候補とし、
語句選択手段が、語句候補抽出手段によって抽出された語句の候補の中から出力対象の語句を選択する
請求項16に記載の情報抽出方法。

- [18] 語句選択手段が、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを語句の候補毎に算出し、語句スコアが所定のしきい値以上である語句の候補を出力対象の語句として選択する
請求項17に記載の情報抽出方法。

- [19] 語句選択手段が、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを語句の候補毎に算出し、語句スコアの高い順に語句の候補をソートし所定の順位以上である語句の候補

を出力対象の語句として選択する

請求項17に記載の情報抽出方法。

- [20] 語句選択手段が、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを語句の候補毎に算出し、語句スコアの高い順に語句の候補をソートし上位の所定の割合の語句の候補を出力対象の語句として選択する

請求項17に記載の情報抽出方法。

- [21] パターン決定手段が、入力語リストに含まれる語句の前方の文字列のみ、あるいは、入力語リストに含まれる語句の後方の文字列のみをパターンとして決定し、

語句候補抽出手段が、前記前方の文字列に続く文字列、あるいは、前記後方の文字列の直前の文字列を文書から抽出して語句の候補とし、

語句選択手段が、語句の候補に含まれる部分文字列を作成し、作成した各部分文字列の中から出力対象の語句を選択する

請求項16に記載の情報抽出方法。

- [22] 語句選択手段が、語句の候補の総数に対する部分文字列を含む語句の候補の割合を、各語句の候補毎および各部分文字列毎に算出して、前記割合が所定の範囲に属する部分文字列を特定し、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを、特定した部分文字列を含む語句の候補毎に算出して、算出結果を前記部分文字列の語句スコアとし、語句スコアが所定のしきい値以上である部分文字列を出力対象の語句として選択する

請求項21に記載の情報抽出方法。

- [23] 語句選択手段が、語句の候補の総数に対する部分文字列を含む語句の候補の割合を、各語句の候補毎および各部分文字列毎に算出して、前記割合が所定の範囲に属する部分文字列を特定し、語句の候補の特定に用いたパターンの候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを、特定した部分文字列を含む語句の候補毎に算出して、算出結果を前記部分文字列の語句スコアとし、語句スコアの高い順に部分文字列をソートし所定の順位以上である部分文

字列を出力対象の語句として選択する

請求項21に記載の情報抽出方法。

- [24] 語句選択手段が、語句の候補の総数に対する部分文字列を含む語句の候補の割合を、各語句の候補毎および各部分文字列毎に算出して、前記割合が所定の範囲に属する部分文字列を特定し、語句の候補の特定に用いたパターン of 候補の重要度の高さ、あるいは語句の候補が出現する文書の多さを示す語句スコアを、特定した部分文字列を含む語句の候補毎に算出して、算出結果を前記部分文字列の語句スコアとし、語句スコアの高い順に部分文字列をソートし上位の所定の割合の部分文字列を出力対象の語句として選択する

請求項21に記載の情報抽出方法。

- [25] 文書検索手段が、形式が異なる文書を含む複数の文書であって、予め記憶装置に記憶されている複数の文書の中から、入力語リストから選択した複数の語句全てを含む文書を検索し、

パターン決定手段が、文書検索手段によって検索された文書毎にパターンの決定を行う

請求項16から請求項24のうちのいずれか1項に記載の情報抽出方法。

- [26] 再作成手段が、語句選択手段によって選択された語句のリストを入力語リストとして文書検索手段に付与する

請求項25に記載の情報抽出方法。

- [27] 再作成手段が、語句選択手段によって選択された語句のリストを入力語リストとしてパターン決定手段に付与する

請求項16から請求項24のうちのいずれか1項に記載の情報抽出方法。

- [28] 文書内の語句を文書から抽出して出力するコンピュータに搭載される情報抽出プログラムであって、

前記コンピュータに、

複数の語句を含む入力語リストを入力する入力処理、

形式が異なる文書を含む複数の文書の中から一つの文書を選択し、選択した文書内で、入力語リストに含まれる語句を抽出対象でない文字列と区切る文字列をパター

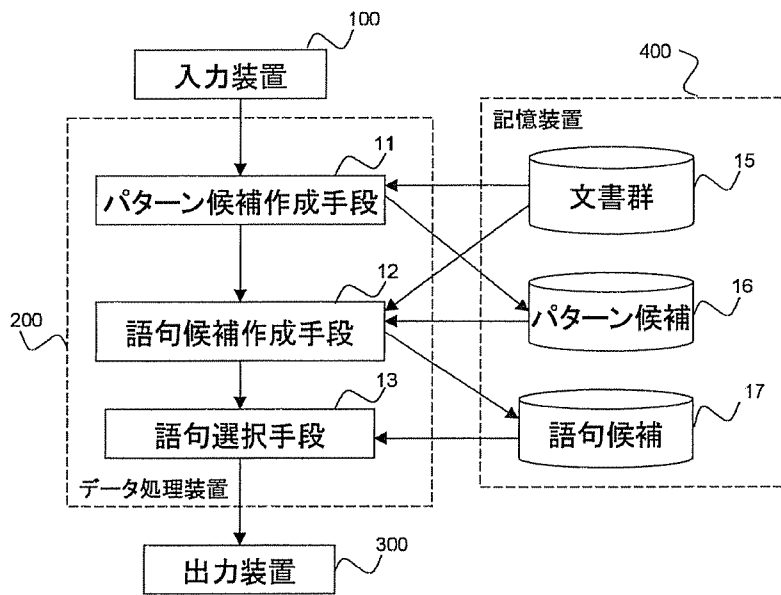
ンとして決定することを文書毎に行うパターン決定処理、

パターンによって区切られる文字列を、当該パターンの決定に用いた文書から抽出して語句の候補とする語句候補抽出処理、および

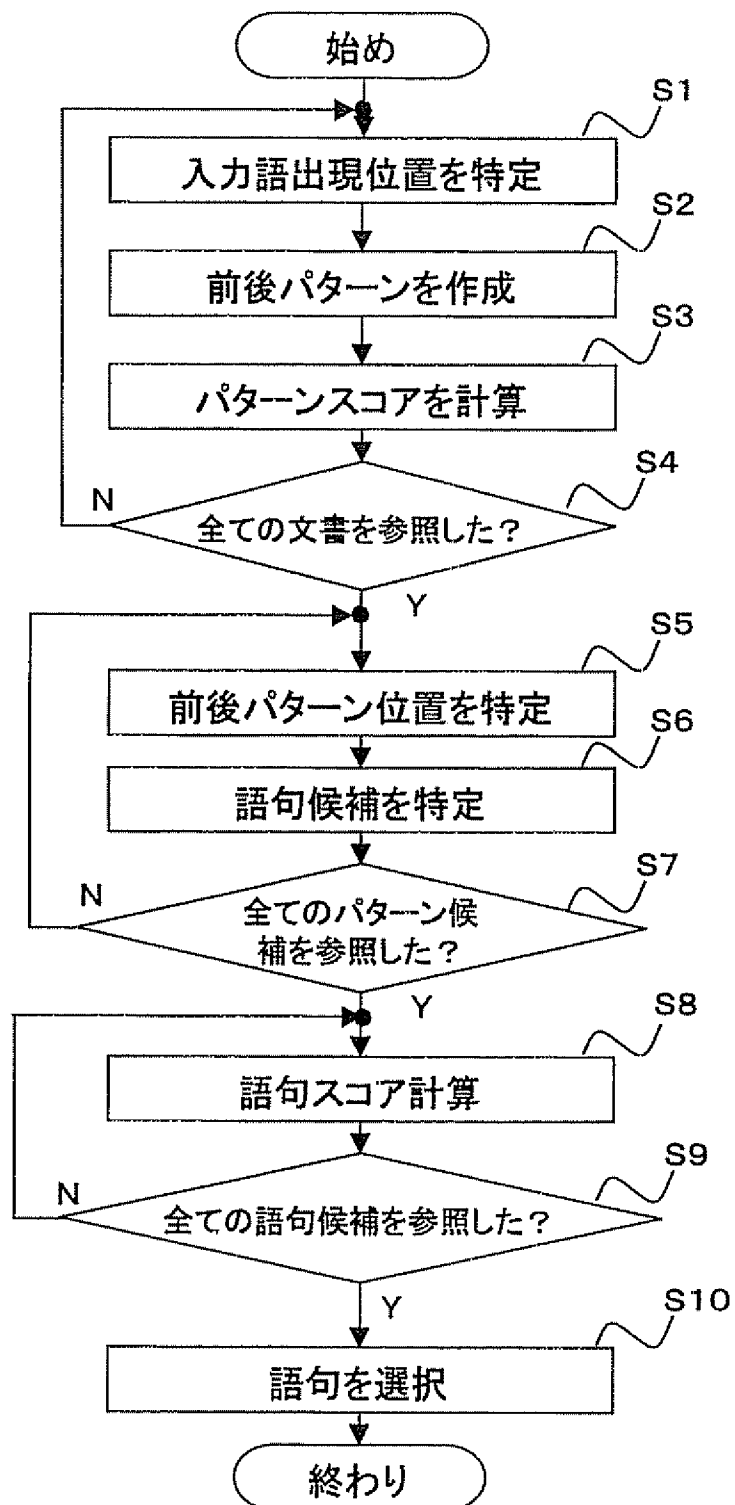
語句候補抽出処理で抽出された語句の候補または当該語句の候補に含まれる部分文字列のうち所定の条件を満たす語句の候補または部分文字列を出力対象の語句として選択する語句選択処理

を実行させるための情報抽出プログラム。

[図1]



[図2]



[図3]

文書ID	文書本文
文書A	<html>...<tr><td>製品A</td>...<tr><td>製品B</td>...<tr><td>製品C</td>...<tr><td>製品D</td>...<tr><td>製品E</td>...
文書B	<html>...製品B 製品A 製品C<p>・製品Cのカテゴリ ...製品D ...
文書C	<html>...<tr><td>企業A</td><td>...
文書D	...商品 製品A 価格 234 商品 製品C 価格 345...
文書E	<html>...<td>製品A</td></tr>...<td>製品B</td></tr>...<td>製品C</td></tr>...
...	...

[図4]

入力語リスト	
製品A	
製品B	
製品C	

[図5]

パターンID	文書ID	前方パターン候補	後方パターン候補	パターンスコア
PAT1	文書A	r><tr><td>	</td><td><	3/3
PAT2	文書B	size="10">	<	3/4
PAT3	文書B	size="10">	<br	2/4
PAT4	文書B	size="10">	<p>	1/4
PAT5	文書B	nt><p>・	のカテゴリ<b	1/4
PAT6	文書D	...	...	...
...	...	...	...	...

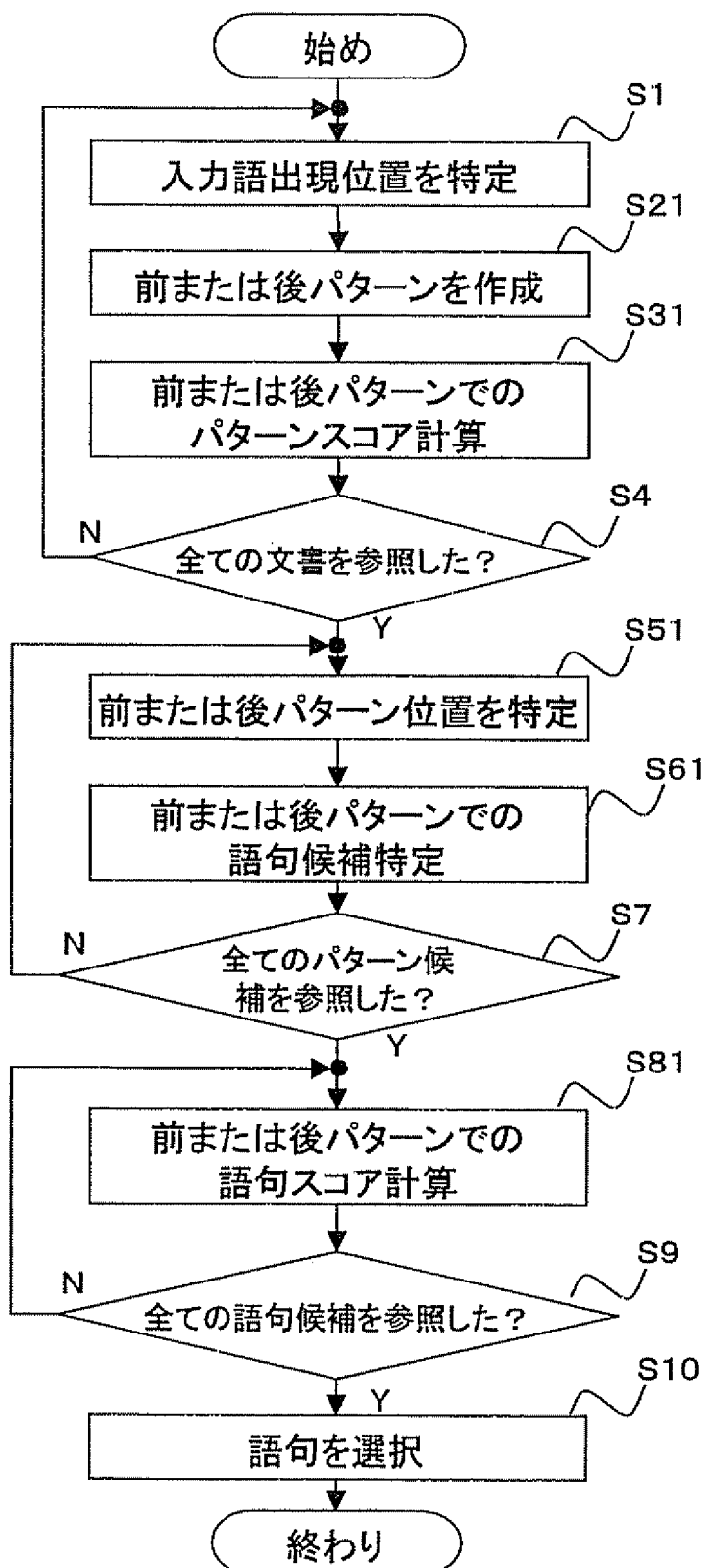
[図6]

文書ID	語句の候補	パターンID	パターンスコア
文書A	製品D	PAT1	3/3
文書A	製品E	PAT1	3/3
文書A	製品F	PAT1	3/3
文書B	製品D	PAT2	3/4
文書B	製品G	PAT2	3/4
文書B	製品D	PAT3	2/4
文書B	製品G	PAT4	1/4
文書B	分類1	PAT5	1/4
文書B	分類2	PAT5	1/4
文書D	製品E	PAT6	2/2
文書D	製品D	PAT6	2/2
文書D	製品G	PAT6	2/2
...	...	...	...

[図7]

出力語	語句スコア
製品D	2.625
製品E	2.000
製品G	2.000
製品F	1.000

[図8]



[図9]

文書ID	文書本文
文書A	<html>...<tr><td>製品A</td>...<tr><td>製品B</td>...<tr><td>製品C</td>...<tr><td>製品D</td>...<tr><td>製品E</td>...
文書B	<html>...製品B 製品A 製品C<p>・製品Cのカテゴリ ...製品D* ...
文書C	<html>...<tr><td>企業A</td><td>...
文書D	...商品 製品A 価格 234 商品 製品C 価格 345...

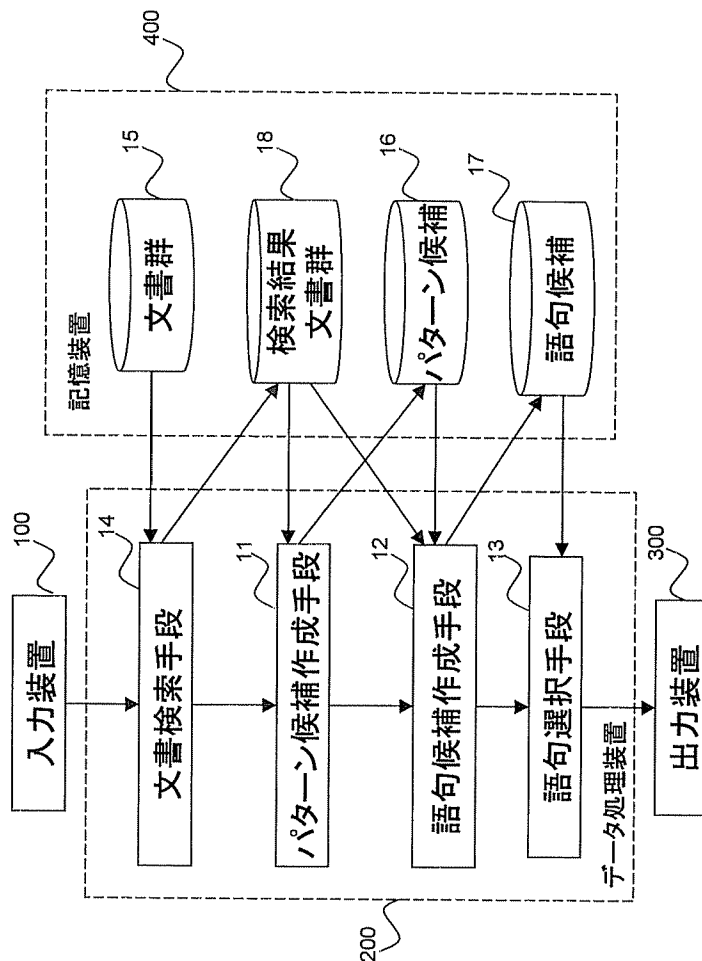
[図10]

パターンID	文書ID	前方パターン候補	パターンスコア
PAT21	文書A	r><tr><td>	3/3
PAT22	文書B	size="10">	3/4
PAT23	文書B	nt><p>・	1/4
PAT24	文書D	価格 234 商品	2/2
...	...	...	...

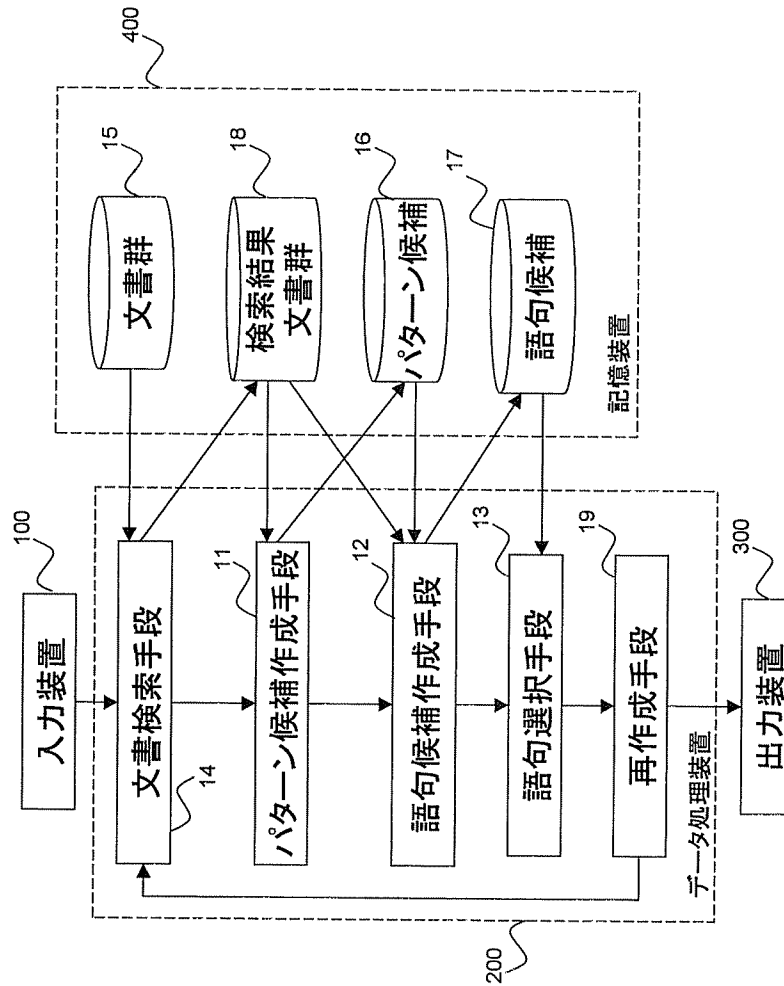
[図11]

文書ID	語句の候補	パターンID	パターンスコア
文書A	製品D</t	PAT21	3/3
文書A	製品E</t	PAT21	3/3
文書A	製品F</t	PAT21	3/3
文書B	製品D*</	PAT22	3/4
文書B	製品G</f	PAT22	3/4
文書B	分類1のカテ	PAT23	1/4
文書B	分類2のカテ	PAT23	1/4
文書D	製品E 価格	PAT24	2/2
文書D	製品D 価格	PAT24	2/2
文書D	製品G 価格	PAT24	2/2

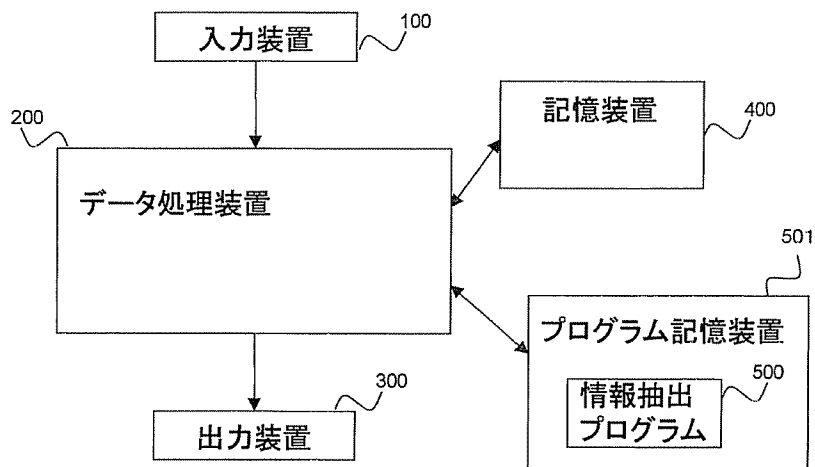
[図12]



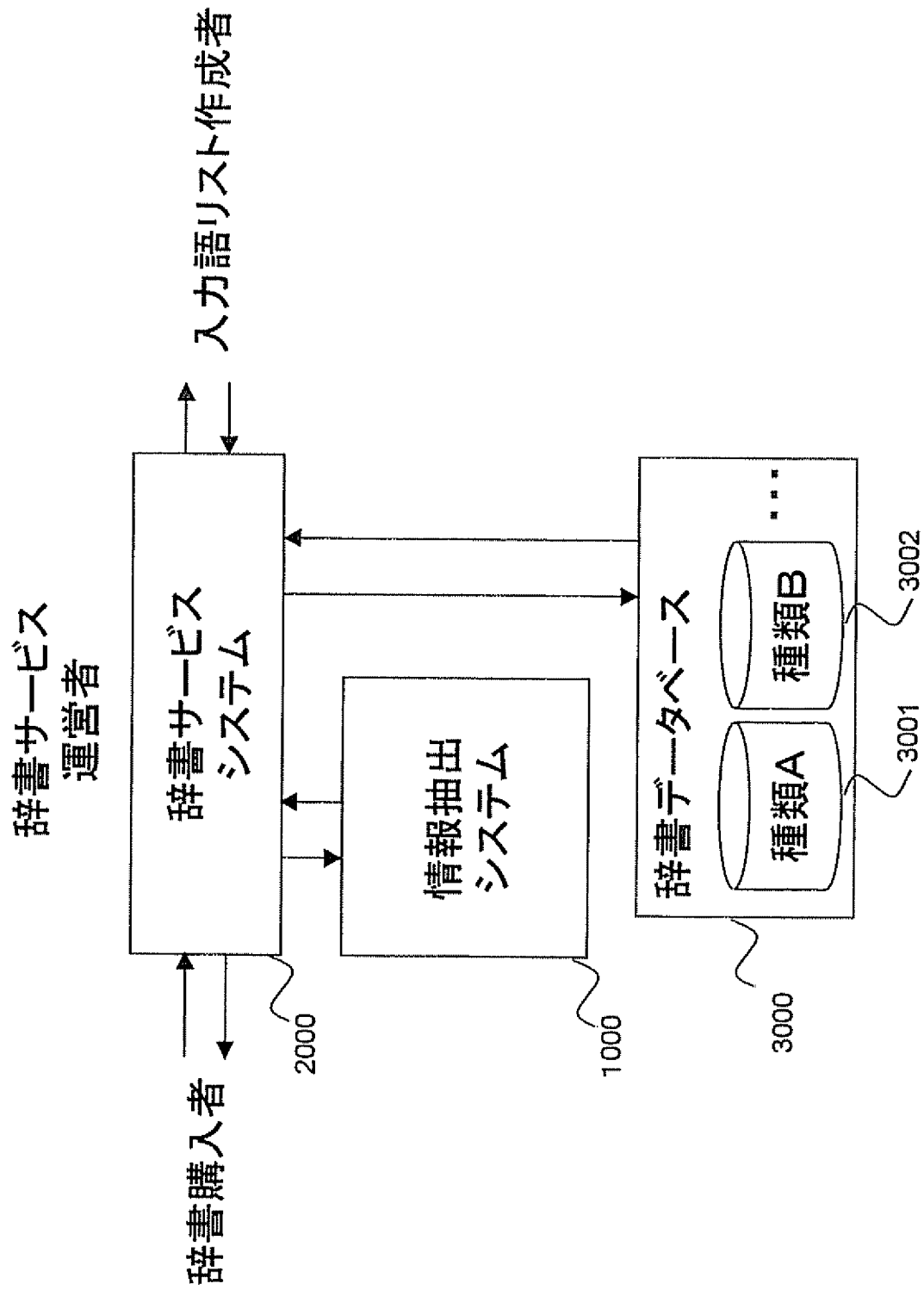
[図13]



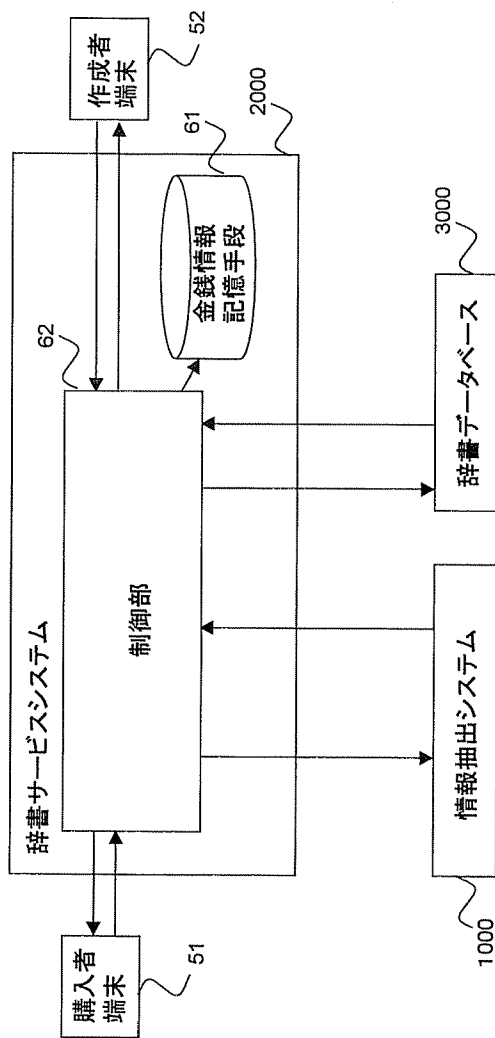
[図14]



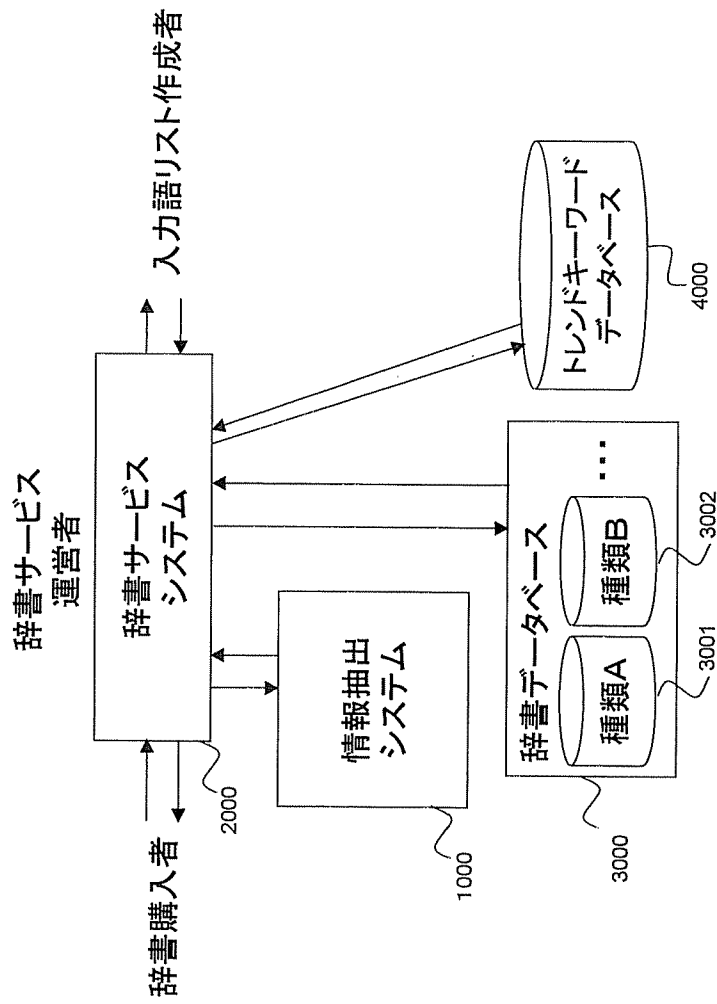
[図15]



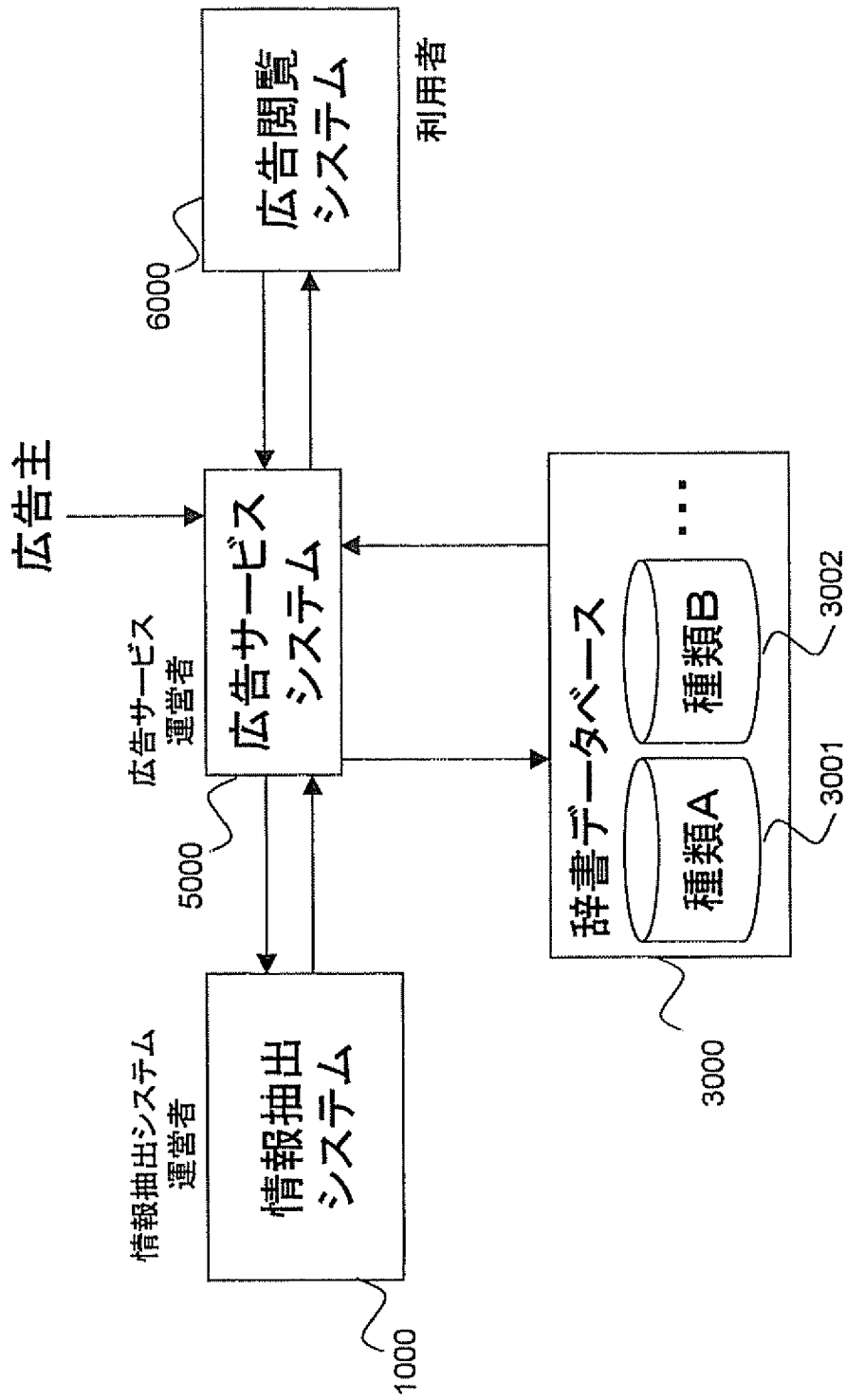
[図16]



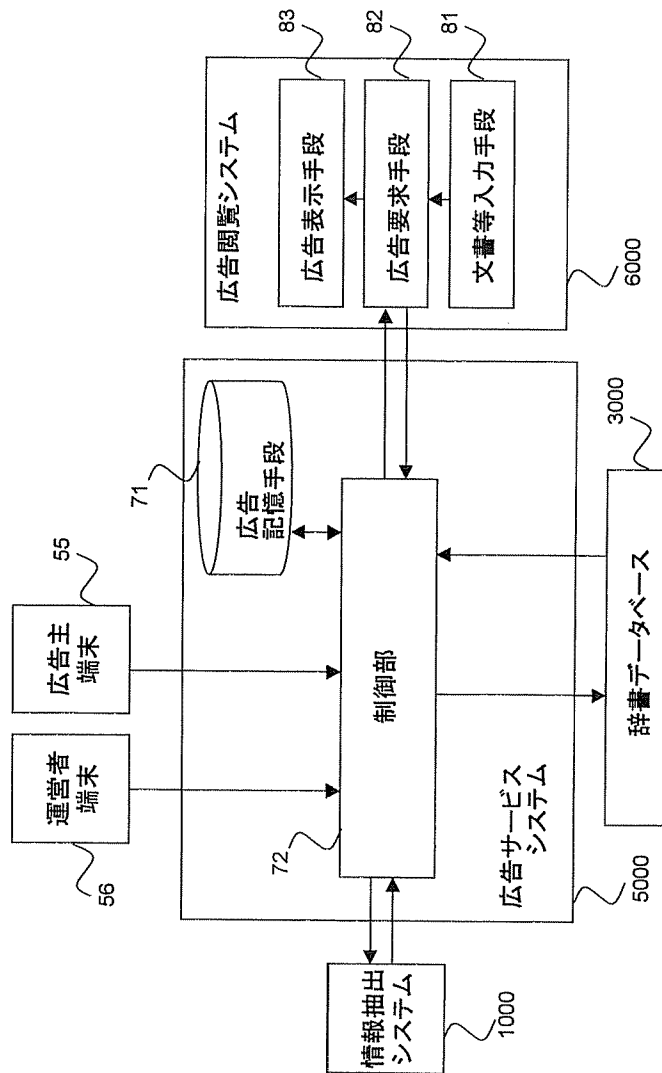
[図17]



[図18]



[図19]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2007/055958

A. CLASSIFICATION OF SUBJECT MATTER

G06F17/30(2006.01) i, G06F17/21(2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F17/30, G06F17/21

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2007
Kokai Jitsuyo Shinan Koho	1971-2007	Toroku Jitsuyo Shinan Koho	1994-2007

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

JSTPlus(JDream2), NRI Cyber Patent

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 2004-151926 A (Ricoh Co., Ltd.), 27 May, 2004 (27.05.04), Full text; Figs. 1 to 24 (Family: none)	1-28
Y	Shinsuke MORI et al., "n-gram Tokei niyoru Corpus Karano Michigo Chushutsu", Transactions of Information Processing Society of Japan, 15 July, 1998 (15.07.98), Vol.39, No.7, pages 2093 to 2100	1-28
A	JP 2002-108887 A (Canon Inc.), 12 April, 2002 (12.04.02), Full text; Figs. 1 to 13 (Family: none)	1-28

☒ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
16 April, 2007 (16.04.07)

Date of mailing of the international search report
24 April, 2007 (24.04.07)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2007/055958

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2001-297100 A (Fuji Xerox Co., Ltd.), 26 October, 2001 (26.10.01), Full text; Figs. 1 to 11 (Family: none)	1-28

A. 発明の属する分野の分類（国際特許分類（I P C））

Int.Cl. G06F17/30(2006.01)i, G06F17/21(2006.01)i

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（I P C））

Int.Cl. G06F17/30, G06F17/21

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922-1996年
日本国公開実用新案公報	1971-2007年
日本国実用新案登録公報	1996-2007年
日本国登録実用新案公報	1994-2007年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JSTPlus(JDream2), NRI サイバーパテント

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
Y	JP 2004-151926 A(株式会社リコー)2004.05.27, 全文, 第1-24図(ファミリーなし)	1-28
Y	森信介、外1名, n グラム統計によるコーパスからの未知語抽出, 情報処理学会論文誌, 1998.07.15, 第39巻, 第7号, p.2093-2100	1-28
A	JP 2002-108887 A(キヤノン株式会社)2002.04.12, 全文, 第1-13図(ファミリーなし)	1-28

☒ C欄の続きにも文献が列挙されている。☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日

16.04.2007

国際調査報告の発送日

24.04.2007

国際調査機関の名称及びあて先

日本国特許庁（I S A/J P）
郵便番号100-8915
東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）

鈴木 和樹

5M

3252

電話番号 03-3581-1101 内線 3599

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	JP 2001-297100 A(富士ゼロックス株式会社)2001. 10. 26, 全文, 第 1-11 図(ファミリーなし)	1-28