

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2010-92350

(P2010-92350A)

(43) 公開日 平成22年4月22日(2010.4.22)

(51) Int.Cl. F I テーマコード (参考)
G 0 6 F 1 7 / 2 1 (2006.01) G O 6 F 1 7 / 2 1 5 O 1 T 5 B 1 O 9
 G O 6 F 1 7 / 2 1 5 5 O A

審査請求 有 請求項の数 7 O L (全 10 頁)

(21) 出願番号	特願2008-262869 (P2008-262869)	(71) 出願人	000004226
(22) 出願日	平成20年10月9日 (2008. 10. 9)		日本電信電話株式会社
			東京都千代田区大手町二丁目3番1号
		(74) 代理人	100069981
			弁理士 吉田 精孝
		(74) 代理人	100087860
			弁理士 長内 行雄
		(74) 代理人	100142789
			弁理士 柳 順一郎
		(72) 発明者	平野 徹
			東京都千代田区大手町二丁目3番1号 日
			本電信電話株式会社内
		(72) 発明者	松尾 義博
			東京都千代田区大手町二丁目3番1号 日
			本電信電話株式会社内
		Fターム(参考)	5B109 NH12 VA02

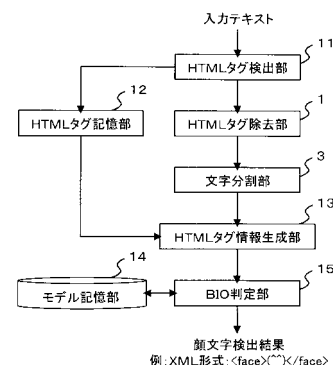
(54) 【発明の名称】 顔文字検出装置、その方法、プログラム及び記録媒体

(57) 【要約】

【課題】テキストからの顔文字検出において、HTMLソーステキストにおける顔文字の出現傾向を捉え、検出精度を向上すること。

【解決手段】HTMLタグ検出部11により、テキスト中のHTMLタグをその出現位置とともに検出してHTMLタグ記憶部12に記憶し、HTMLタグが除去され、文字単位に分割された前記テキスト中の各文字の直前および直後にHTMLタグがあったかどうかを表すHTMLタグ情報を、HTMLタグ情報生成部13により、HTMLタグ記憶部12に記憶されたHTMLタグとその出現位置に基づいて生成し、これを当該テキスト中の各文字の表記とともに素性として用いて、BIO判定部15により、モデル記憶部14に記憶されたモデルを用いて顔文字を構成する最初の文字“B”、顔文字を構成する2番目以降の文字“I”、顔文字以外の文字“O”のいずれに当たるかを機械学習で判定する。

【選択図】図6



【特許請求の範囲】**【請求項 1】**

HTML タグを含むテキスト中の顔文字を検出する装置であって、
テキスト中の HTML タグをその出現位置とともに検出する HTML タグ検出部と、
検出された HTML タグを前記テキストにおけるその出現位置とともに記憶する HTML
タグ記憶部と、
前記テキストから HTML タグを除去する HTML タグ除去部と、
HTML タグが除去された前記テキストを文字単位に分割する文字分割部と、
HTML タグが除去され、文字単位に分割された前記テキスト中の各文字の直前および
直後に HTML タグがあったかどうか、あった場合はどのような HTML タグかを表す H
TML タグ情報を、前記 HTML タグ記憶部に記憶された前記テキスト中の HTML タグ
とその出現位置に基づいて生成する HTML タグ情報生成部と、
少なくともテキスト中の各文字の表記並びにその直前および直後の HTML タグ情報を
素性として、当該テキスト中の各文字が顔文字を構成する文字列の最初の文字を意味する
“ B ”、顔文字を構成する文字列の 2 番目以降の文字を意味する “ I ”、顔文字以外の文
字を意味する “ O ” のいずれに当たるかを判定するためのモデルを記憶するモデル記憶部
と、
HTML タグが除去され、文字単位に分割された前記テキスト中の各文字の表記並びに
その直前および直後の HTML タグ情報を入力とし、前記モデル記憶部に記憶されたモデ
ルを用いて、前記テキスト中の各文字が前記 “ B ”、“ I ”、“ O ” のいずれに当たるか
を判定する B I O 判定部とを備えた
ことを特徴とする顔文字検出装置。

【請求項 2】

HTML タグ情報生成は、
HTML タグが除去され、文字単位に分割された前記テキスト中の各文字の直前および
直後の位置と、前記 HTML タグ記憶部に記憶された前記テキストにおける HTML タグ
の出現位置とを比較し、一致する場合はその HTML タグを HTML タグ情報として、ま
た、一致しない場合はタグ無しを意味する HTML タグを HTML タグ情報として、当該
テキスト中の各文字の表記に対応させて出力することで行う
ことを特徴とする請求項 1 に記載の顔文字検出装置。

【請求項 3】

HTML タグを含むテキスト中の顔文字を検出する方法であって、
HTML タグ検出部が、テキスト中の HTML タグをその出現位置とともに検出し、H
TML タグ記憶部に記憶するステップと、
HTML タグ除去部が、前記テキストから HTML タグを除去するステップと、
文字分割部が、HTML タグが除去された前記テキストを文字単位に分割するステッ
プと、
HTML タグ情報生成部が、HTML タグが除去され、文字単位に分割された前記テキ
スト中の各文字の直前および直後に HTML タグがあったかどうか、あった場合はどのよ
うな HTML タグかを表す HTML タグ情報を、前記 HTML タグ記憶部に記憶された前
記テキスト中の HTML タグとその出現位置に基づいて生成するステップと、
B I O 判定部が、HTML タグが除去され、文字単位に分割された前記テキスト中の各
文字の表記並びにその直前および直後の HTML タグ情報を入力とし、少なくともテキ
スト中の各文字の表記並びにその直前および直後の HTML タグ情報を素性として、当該
テキスト中の各文字が顔文字を構成する文字列の最初の文字を意味する “ B ”、顔文字を構
成する文字列の 2 番目以降の文字を意味する “ I ”、顔文字以外の文字を意味する “ O ”
のいずれに当たるかを判定するためのモデルを用いて、前記テキスト中の各文字が前記 “
B ”、“ I ”、“ O ” のいずれに当たるかを判定するステップとを含む
ことを特徴とする顔文字検出方法。

【請求項 4】

10

20

30

40

50

HTMLタグ情報生成は、

HTMLタグが除去され、文字単位に分割された前記テキスト中の各文字の直前および直後の位置と、前記HTMLタグ記憶部に記憶された前記テキストにおけるHTMLタグの出現位置とを比較し、一致する場合はそのHTMLタグをHTMLタグ情報として、また、一致しない場合はタグ無しを意味するHTMLタグをHTMLタグ情報として、当該テキスト中の各文字の表記に対応させて出力することで行う

ことを特徴とする請求項3に記載の顔文字検出方法。

【請求項5】

コンピュータを、請求項1または2に記載の顔文字検出装置の各手段として機能させるためのプログラム。

10

【請求項6】

コンピュータに、請求項3または4に記載の顔文字検出方法の各処理ステップを実行させるためのプログラム。

【請求項7】

請求項5または6に記載のプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、HTMLソーステキスト等のHTMLタグを含むテキスト中の顔文字を検出する技術に関する。

20

【背景技術】

【0002】

顔文字とは、テキスト中で使用可能な文字を組み合わせる様々な感情等に対応する擬似的な人の表情を表すようにしたもので、パーソナルコンピュータや携帯電話における電子メールなどのテキストベースのコミュニケーションツールにおいて、書き手（送信側）の感情等を読み手（受信側）に伝える手段として用いられることが多い。例えば、「今日はとっても楽しかったです。(^^)明日も楽しみ！」というテキスト中の文字列「(^^)」が顔文字であり、この場合、「喜び」の感情を表している。

【0003】

30

従来、テキスト中の顔文字を検出する技術としては、既存のHTMLタグ除去技術、形態素解析技術を利用してテキスト中の各文字が属する形態素の品詞を抽出し、これを当該テキスト中の各文字の表記とともに素性として用いて、顔文字を構成する文字列の最初の文字を意味する“B”、顔文字を構成する文字列の2番目以降の文字を意味する“I”、顔文字以外の文字を意味する“O”（以下、これらを顔文字タグと呼ぶ。）のいずれに当たるかを機械学習で推定（判定）する方法があった（非特許文献1参照）。例えば、「今日はとっても楽しかったです。(^^)
明日も楽しみ！
」というテキストにおいて、文字「(」が顔文字タグ“B”であることを推定する場合は、図1に示すような太線内の情報を素性として利用していた。

【0004】

40

図2は前述した従来の顔文字検出技術にかかる装置の一例を示すもので、HTMLタグ除去部1、形態素解析部2、文字分割部3、形態素品詞抽出部4、モデル記憶部5およびBIO判定部6からなる。

【0005】

HTMLタグ除去部1では、既存のHTMLタグ除去技術により、入力されたテキストからHTMLタグを除去する。例えば、入力されたテキストが「今日はとっても楽しかったです。(^^)
明日も楽しみ！
」の場合、HTMLタグ“
”が除かれ、「今日はとっても楽しかったです。(^^)明日も楽しみ！」となる。

【0006】

形態素解析部2では、既存の形態素解析技術により、HTMLタグ除去部1でHTML

50

タグが除去されたテキストを単語に区切り、さらに各単語に品詞を付与する。例えば、HTMLタグ除去後のテキストが「今日はとっても楽しかったです。(^^)明日も楽しみ!」の場合、区切り記号を“/”、品詞を“[品詞]”で表すと、「今日[名詞]/は[助詞]/とっても[副詞]/楽し[形容詞]/かった[形容詞接尾辞]/です[助動詞]/。[記号-句点]/([記号-括弧開]/^[記号-その他]/^[記号-その他]/)[記号-括弧閉]/明日[名詞]/も[助詞]/楽しみ[名詞]/! [記号-その他]」となる。

【0007】

文字分割部3では、HTMLタグ除去部1でHTMLタグが除去されたテキストを文字単位(文字コード単位)に分割する。例えば、前記同様、HTMLタグ除去後のテキストが「今日はとっても楽しかったです。(^^)明日も楽しみ!」の場合、区切り記号を“/”

10

【0008】

形態素品詞抽出部4では、形態素解析部2での結果と文字分割部3での結果とを入力とし、入力されたテキスト中の各文字の表記および各文字が属する形態素の品詞を抽出する。例えば、形態素解析部2での結果が「今日[名詞]/は[助詞]/とっても[副詞]/楽し[形容詞]/かった[形容詞接尾辞]/です[助動詞]/。[記号-句点]/([記号-括弧開]/^[記号-その他]/^[記号-その他]/)[記号-括弧閉]/明日[名詞]/も[助詞]/楽しみ[名詞]/! [記号-その他]」であり、文字分割部3での結果が「今/日/は/と/っ/て/も/楽/し/か/っ/た/で/す/。/(/^(^)/明/日/も/楽/し/み/!」である場合、図3に

20

【0009】

モデル記憶部5では、予め機械学習により生成した、少なくともテキスト中の各文字の表記および各文字が属する形態素の品詞を素性、詳細には図1に示したようにテキスト中の各文字のその文字を含む前後n(ここではn=3)文字の表記、それらの形態素の品詞および前n文字についてのBIO判定結果を素性として、当該テキスト中の各文字が顔文字タグ“B”、“I”、“O”のいずれに当たるかを判定するためのモデルを記憶している。但し、使用する機械学習によっては、前n文字についてのBIO判定結果を明示的に与える必要はない。

30

【0010】

BIO判定部6では、形態素品詞抽出部4での結果を入力とし、モデル記憶部5に記憶されたモデルを用いて、テキスト中の各文字について、その先頭から順に顔文字タグ“B”、“I”、“O”のいずれに当たるかを判定し、顔文字タグ“B”および“I”に対応する文字列を検出結果として出力する。例えば、図3に示した結果が入力された場合、図4に示すように、先頭から順に、各文字のその文字を含む前後n(ここではn=3)文字の表記、それらの形態素の品詞および前n文字についてのBIO判定結果を素性としてモデル記憶部5に記憶されたモデルに入力してBIO判定結果を得て、顔文字タグ“B”および“I”と判定された文字列「(^^)」を、例えばXML形式「<face>(^^)</face>」で出力する。

40

【0011】

モデル記憶部5に記憶するモデルは、テキスト集合の各テキストに対し、前記HTMLタグ除去部1、形態素解析部2、文字分割部3、形態素品詞抽出部4での処理を行い、その結果として得られる、HTMLタグが除去され、文字単位に分割されたテキスト中の各文字の表記および各文字が属する形態素の品詞の情報と、前記テキスト中の各文字について前記顔文字タグ“B”、“I”、“O”のいずれかに当たるかを人手により判断した結果としての情報とを用い、既存の機械学習によって予め作成したものである。

【0012】

この際、使用する機械学習としてはどのようなものでも良いが、系列ラベリング問題のためのものが望ましい。系列ラベリング問題とは、与えられた入力トークン列 $x = (x_1, x_2, \dots, x_N)$ に対して、適切な出力ラベル列 $y = (y_1, y_2, \dots, y_N)$ を対応づけ

50

る問題である。

【非特許文献 1】田中裕紀、高村大也、奥村学、「文字ベースのコミュニケーションにおける顔文字に関する研究」、言語処理学会第 10 回年次大会、D4-3、2004

【発明の開示】

【発明が解決しようとする課題】

【0013】

ところで、HTML ソーステキストの場合、顔文字は“
”などの HTML タグの直前に出現することが多い。例えば、「今日はとっても楽しかったです。(^^)
」というテキストでは、顔文字「(^^)」は HTML タグ“
”の直前に出現している。

【0014】

前述した従来の顔文字検出技術では、入力されたテキストに HTML タグが含まれている場合、前処理で HTML タグを除去し、その情報を全く利用していないため、適切に顔文字を検出できないという問題があった。

【0015】

本発明の目的は、テキストからの顔文字検出において、HTML ソーステキストにおける顔文字の出現傾向を捉え、検出精度を向上することにある。

【課題を解決するための手段】

【0016】

本発明では、前記目的を達成するため、HTML タグを含むテキスト中の HTML タグをその出現位置とともに検出して記憶しておき、一方、前記テキストから HTML タグを除去するとともに文字単位に分割し、該 HTML タグが除去され、文字単位に分割された前記テキスト中の各文字の直前および直後に HTML タグがあったかどうか、あった場合はどのような HTML タグかを表す HTML タグ情報を、前記記憶した前記テキストにおける HTML タグの出現位置に基づいて生成し、これを当該テキスト中の各文字の表記とともに素性として用いて、顔文字を構成する文字列の最初の文字を意味する“B”、顔文字を構成する文字列の 2 番目以降の文字を意味する“I”、顔文字以外の文字を意味する“O”のいずれに当たるかを機械学習で推定（判定）することを特徴とする。例えば、「今日はとっても楽しかったです。(^^)
明日も楽しみ！
」というテキストにおいて、文字「(」が顔文字タグ“B”であることを推定する場合は、図 5 に示すような太線内の情報を素性として利用する。

【発明の効果】

【0017】

本発明によれば、テキスト中の各文字の表記とともにその直前および直後に HTML タグがあったかどうかを表す HTML タグ情報を素性として機械学習で推定（判定）するため、HTML ソーステキストにおける顔文字の出現傾向を捉え、検出精度を向上することが可能となる。

【発明を実施するための最良の形態】

【0018】

図 6 は本発明の顔文字検出装置の実施の形態の一例を示すもので、図中、従来例と同一構成部分は同一符号をもって表す。即ち、1 は HTML タグ除去部、3 は文字分割部、11 は HTML タグ検出部、12 は HTML タグ記憶部、13 は HTML タグ情報生成部、14 はモデル記憶部、15 は B I O 判定部である。

【0019】

HTML タグ検出部 11 は、HTML タグを含むテキストから HTML タグを検出し、該検出した HTML タグの先頭の文字位置を出現位置として取得する。ここで、HTML タグを検出する具体的な方法としては、テキスト中で使用される HTML タグの集合を格納したリストを予め用意しておき、入力されたテキストに対してリスト内の HTML タグ毎にパターンマッチングを行うことでできる。また、入力されたテキストはそのまま HTML タグ除去部 1 へ出力される。

【0020】

10

20

30

40

50

HTMLタグ記憶部12は、HTMLタグ検出部11で検出されたHTMLタグを前記テキストにおけるその出現位置とともに記憶する。また、この際、後述するHTMLタグ除去後のテキストにおける文字位置との整合を図るため、検出したHTMLタグの文字数を併せて記憶しておくものとする。なお、HTMLタグの文字数は、パターンマッチングの際にHTMLタグと一致した文字列の先頭および末尾の文字位置の差から求める外、前述したHTMLタグのリストにおいて予め各HTMLタグに対応づけてその文字数を格納しておき、ここから読み出すようにしても良い。

【0021】

HTMLタグ情報生成部13は、HTMLタグ除去部1でHTMLタグが除去され、文字分割部3で文字単位（文字コード単位）に分割されたテキストと、HTMLタグ記憶部12に記憶されたテキスト中のHTMLタグ、その出現位置および文字数とを入力として、HTMLタグが除去され、文字単位に分割されたテキスト中の各文字の直前および直後にHTMLタグがあったかどうかを判定し、あった場合はそのHTMLタグを、また、なかった場合は無しを意味するHTMLタグ（ここではNULL）をHTMLタグ情報として、当該テキスト中の各文字の表記に対応させて出力する。

【0022】

詳細には、HTMLタグが除去され、文字単位に分割されたテキスト中の一の文字の直前の文字位置および直後の文字位置、実際には当該文字の文字位置およびこれに1を加えた文字位置と、前記HTMLタグ記憶部12に記憶された前記テキスト中の各HTMLタグの出現位置（タグの先頭の文字位置）とを順次比較し、一致する場合はそのHTMLタグをHTMLタグ情報として出力し、また、一致しない場合はタグ無しを意味するHTMLタグ（NULL）をHTMLタグ情報として出力し、これをテキスト中の全ての文字について繰り返し行う。但し、この際、テキスト中で最初に出現したHTMLタグの出現位置はHTMLタグ記憶部12に記憶された値そのものを用いるが、2つ目以降に出現したHTMLタグの出現位置については、HTMLタグ記憶部12に記憶された値から、それ以前に出現したHTMLタグの文字数の和を差し引いた値を用いるものとする。

【0023】

モデル記憶部14は、予め機械学習により生成した、少なくともテキスト中の各文字の表記並びにその直前および直後のHTMLタグ情報を素性、詳細には図5に示したようにテキスト中の各文字のその文字を含む前後 n （ここでは $n=3$ ）文字の表記、それらの直前および直後のHTMLタグ情報、更には前 n 文字についてのBIO判定結果を素性として、当該テキスト中の各文字が顔文字タグ“B”、“I”、“O”のいずれに当たるかを判定するためのモデルを記憶している。但し、使用する機械学習によっては、前 n 文字についてのBIO判定結果を明示的に与える必要はない。

【0024】

BIO判定部15は、HTMLタグ情報生成部13での結果を入力とし、モデル記憶部14に記憶されたモデルを用いて、テキスト中の各文字について、その先頭から順に顔文字タグ“B”、“I”、“O”のいずれに当たるかを判定し、顔文字タグ“B”および“I”に対応する文字列を検出結果として出力、例えば顔文字の開始タグ“<face>”と顔文字の終了タグ“</face>”との間に顔文字タグ“B”および“I”に対応する文字列「 $x \ x \ x$ 」を記述したXML形式「<face> $x \ x \ x \ x$ </face>」で出力する。

【0025】

モデル記憶部14に記憶するモデルは、テキスト集合の各テキストに対し、前記HTMLタグ検出部11、HTMLタグ記憶部12、HTMLタグ除去部1、文字分割部3、HTMLタグ情報生成部13での処理を行い、その結果として得られる、HTMLタグが除去され、文字単位に分割されたテキスト中の各文字の表記並びにその直前および直後のHTMLタグ情報と、前記テキスト中の各文字について前記顔文字タグ“B”、“I”、“O”のいずれかに当たるかを人手により判断した結果としての情報とを用い、既存の機械学習によって予め作成したものである。この際、使用する機械学習としてはどのようなものでも良いが、系列ラベリング問題のためのものが望ましい。

10

20

30

40

50

【 0 0 2 6 】

図 7 は本発明の顔文字検出装置における処理の流れを示すもので、以下、例を挙げてその動作を詳細に説明する。

【 0 0 2 7 】

HTML タグ検出部 1 1 に入力されたテキスト (HTML タグを含むテキスト) は、当該テキスト中の HTML タグが検出され、その出現位置および文字数が該検出された HTML タグとともに HTML タグ記憶部 1 2 に出力されて記憶される (s 1)。例えば、入力テキストが「今日はとっても楽しかったです。(^^)
明日も楽しみ!
」であれば、1 つ目の HTML タグ "
" の先頭の文字位置は「20」であり、2 つ目の HTML タグ "
" の先頭の文字位置は「31」であり、文字数はいずれも「4」であるから、図 8 に示すような結果が得られる。なお、図 8 中の「タグ番号」とは、後述する説明のために出現順に単純に付与した、テキスト中の HTML タグの識別番号である。また、入力されたテキストはそのまま HTML タグ除去部 1 へ出力される。

10

【 0 0 2 8 】

HTML タグ除去部 1 に入力されたテキストは、既存の HTML タグ除去技術により、当該テキスト中の HTML タグが除去され、文字分割部 3 に出力される (s 2)。例えば、入力テキストが「今日はとっても楽しかったです。(^^)
明日も楽しみ!
」であれば、HTML タグ "
" が除かれ、「今日はとっても楽しかったです。(^^)明日も楽しみ!」となる。

20

【 0 0 2 9 】

文字分割部 3 では、HTML タグ除去部 1 で HTML タグが除去されたテキストを文字単位 (文字コード単位) に分割し、HTML タグ情報生成部 1 3 に出力する (s 3)。例えば、HTML タグ除去後のテキストが「今日はとっても楽しかったです。(^^)明日も楽しみ!」の場合、区切り記号を "/" で表すと、「今/日/は/と/っ/て/も/楽/し/か/っ/た/で/す/。(^^)/明/日/も/楽/し/み/!」となる。

【 0 0 3 0 】

HTML タグ情報生成部 1 3 では、HTML タグ記憶部 1 2 に記憶されたテキスト中の HTML タグ、その出現位置および文字数に基づいて、HTML タグ除去部 1 で HTML タグが除去され、文字分割部 3 で文字単位に分割されたテキスト中の各文字について、その直前および直後に HTML タグがあったかどうかを判定し、あった場合はその HTML タグを、また、なかった場合は無しを意味する HTML タグ (NULL) を HTML タグ情報として、当該テキスト中の各文字の表記に対応させて BIO 判定部 1 5 に出力する (s 4)。

30

【 0 0 3 1 】

例えば、HTML タグ記憶部 1 2 に記憶されたテキスト中の HTML タグ、その出現位置および文字数が図 8 に示したものであり、HTML タグが除去され、文字単位に分割されたテキストが「今/日/は/と/っ/て/も/楽/し/か/っ/た/で/す/。(^^)/明/日/も/楽/し/み/!」であった場合、図 9 に示すような結果が得られる。

【 0 0 3 2 】

即ち、文字位置「19」の文字「)」の直後の文字位置、つまり「19」に 1 を加えた文字位置「20」と、タグ番号 1 の HTML タグの出現位置「20」とが一致するため、文字位置「19」の文字「)」の直後の HTML タグ情報は、当該タグ番号 1 の HTML タグ "
" となる。同様に、文字位置「20」の文字「明」の直前の文字位置、つまり当該文字「明」の文字位置「20」と、タグ番号 1 の HTML タグの出現位置「20」とが一致するため、文字位置「20」の文字「明」の直前の HTML タグ情報も、当該タグ番号 1 の HTML タグ "
" となる。一方、文字位置「26」の文字「!」の直後の文字位置、つまり「26」に 1 を加えた文字位置「27」と、タグ番号 2 の HTML タグの出現位置、ここでは「31」からそれ以前に出現したタグ番号 1 の HTML タグの文字数「4」を差し引いた値である「27」とが一致するため、文字位置「26」の文字「!」の直後の HTML タグ情報は、当該タグ番号 2 の HTML タグ "
" となる。

40

50

【 0 0 3 3 】

B I O 判定部 1 5 では、H T M L タグ情報生成部 1 3 での結果を入力とし、モデル記憶部 1 4 に記憶されたモデルを用いて、テキスト中の各文字について、その先頭から順に顔文字タグ “ B ”、“ I ”、“ O ”のいずれに当たるかを判定し、顔文字タグ “ B ”および “ I ”に対応する文字列を検出結果として出力する (s 5)。例えば、図 9 に示した結果が入力された場合、図 1 0 に示すように、先頭から順に、各文字のその文字を含む前後 n (ここでは n = 3) 文字の表記、それらの直前および直後の H T M L タグ情報、更には前 n 文字についての B I O 判定結果を素性としてモデル記憶部 1 4 に記憶されたモデルに入力して B I O 判定結果を得て、顔文字タグ “ B ”および “ I ”と判定された文字列 「 (^ ^) 」を X M L 形式 「 < face> (^ ^)< / face> 」で出力する。

10

【 0 0 3 4 】

なお、本発明は、周知のコンピュータに媒体もしくは通信回線を介して、図 6 の構成図に示された機能を実現するプログラムあるいは図 7 のフローチャートに示された手順を備えるプログラムをインストールすることによっても実現可能である。

【 産業上の利用可能性 】

【 0 0 3 5 】

本発明によってテキスト中の顔文字を検出し、これを削除することで、当該テキストに対する形態素解析の精度を向上させることが可能となる。

【 図面の簡単な説明 】

【 0 0 3 6 】

20

【 図 1 】 従来の顔文字検出技術における素性の説明図

【 図 2 】 従来の顔文字検出装置の一例を示す構成図

【 図 3 】 形態素品詞抽出結果の一例を示す説明図

【 図 4 】 従来の顔文字検出技術による B I O 判定のようすを示す説明図

【 図 5 】 本発明の顔文字検出技術における素性の説明図

【 図 6 】 本発明の顔文字検出装置の実施の形態の一例を示す構成図

【 図 7 】 本発明の顔文字検出装置における処理の流れを示すフローチャート

【 図 8 】 H T M L タグ検出結果の一例を示す説明図

【 図 9 】 H T M L タグ情報生成結果の一例を示す説明図

【 図 1 0 】 本発明の顔文字検出技術による B I O 判定のようすを示す説明図

30

【 符号の説明 】

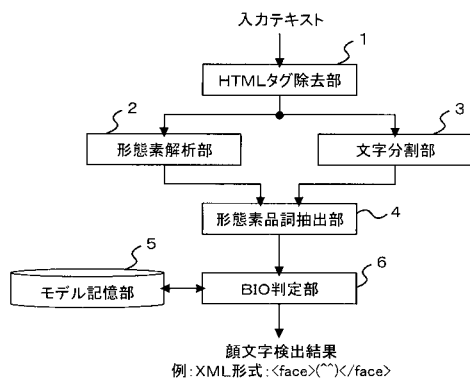
【 0 0 3 7 】

1 : H T M L タグ除去部、 3 : 文字分割部、 1 1 : H T M L タグ検出部、 1 2 : H T M L タグ記憶部、 1 3 : H T M L タグ情報生成部、 1 4 : モデル記憶部、 1 5 : B I O 判定部。

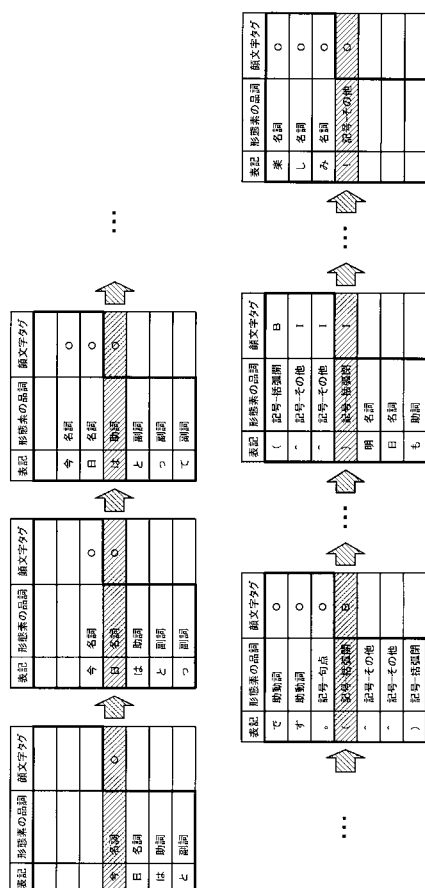
【図 1】

表記	形態素の品詞	顔文字タグ
で	助動詞	○
す	助動詞	○
。	記号-句点	○
(	記号-括弧開	B
^	記号-その他	
^	記号-その他	
)	記号-括弧閉	

【図 2】



【図 4】



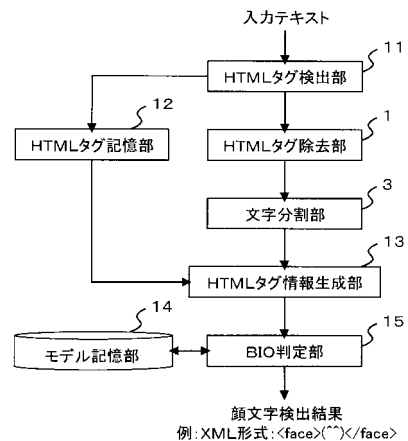
【図 3】

表記	形態素の品詞	表記	形態素の品詞	表記	形態素の品詞
今	名詞	か	形容詞接尾辞	)	記号-括弧閉
日	名詞	っ	形容詞接尾辞	明	名詞
は	助詞	た	形容詞接尾辞	日	名詞
と	副詞	で	助動詞	も	助詞
っ	副詞	す	助動詞	楽	名詞
て	副詞	。	記号-句点	し	名詞
も	副詞	(	記号-括弧開	み	名詞
楽	形容詞	^	記号-その他	!	記号-その他
し	形容詞	^	記号-その他		

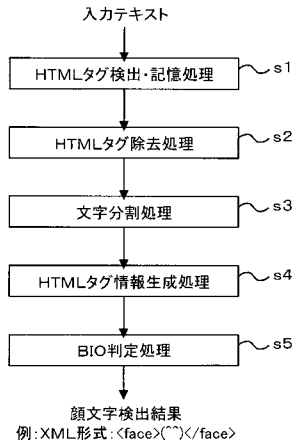
【図 5】

表記	直前のHTMLタグ	直後のHTMLタグ	顔文字タグ
で	NULL	NULL	○
す	NULL	NULL	○
。	NULL	NULL	○
(	NULL	NULL	B
^	NULL	NULL	
^	NULL	NULL	
)	NULL	 	

【図 6】



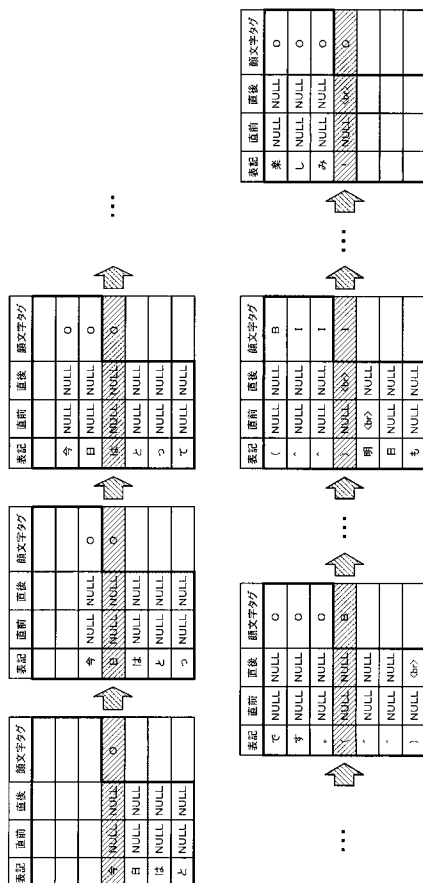
【 図 7 】



【圖 8】

タグ番号	出現位置	HTMLタグ	文字数
1	20	 	4
2	31	 	4

【 図 1 0 】



【 図 9 】

文字位置	表記	直前	直後	文字位置	表記	直前	直後
1	今	NULL	NULL	14	ず	NULL	NULL
2	日	NULL	NULL	15	。	NULL	NULL
3	は	NULL	NULL	16	(	NULL	NULL
4	と	NULL	NULL	17	^	NULL	NULL
5	っ	NULL	NULL	18	^	NULL	NULL
6	て	NULL	NULL	19	)	NULL	
7	も	NULL	NULL	20	明	 	NULL
8	楽	NULL	NULL	21	日	NULL	NULL
9	し	NULL	NULL	22	も	NULL	NULL
10	か	NULL	NULL	23	楽	NULL	NULL
11	っ	NULL	NULL	24	し	NULL	NULL
12	た	NULL	NULL	25	み	NULL	NULL
13	で	NULL	NULL	26	!	NULL	