社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－GN－68（1）
2008/5/23

webコンテンツを直接入力可能な日本語変換システム

小林孝典　市村哲

東京工科大学大学院

文章作成時，webの情報や画像を参照したいときwebブラウザを開いて探し出すのには多くの手間がかかる．例えば画像を貼り付ける時，画像の素材を探し出して文章内に貼り付けるまでにいくつかの手順が必要である．これらには，「文章の内容にあったwebコンテンツを自動的に探してくる」ことと，探してきた「webコンテンツを使用したいときすぐに持って来る」手順が含まれる．

そこで本論文では，日本語変換時にwebから入力文字に関連するwebコンテンツをダウンロードし，その後の変換候補選択時にダウンロードしたコンテンツを表示できるシステム「IMEdia」を提案する．　実験の結果，ブラウザを開かずに画像を取得することにより，スピードがあがった，画像が色々でてくるので楽しい，等の評価が得られた．また，表示される画像は作文時のイメージを膨らませる発想支援としても役立つことがわかった．

A Kanji conversion system that can diretly input web contents

Takanori Kobayashi , Satoshi Ichimura

Tokyo University of Technology

It takes a lot of time to open or search web pages.For instance, many procedures are necessary when searching the material of the image, and putting it into the document. In this paper we proposes a kanji conversion system that can directly input web contents. into the document.As a result of the experiment, we have obained comments from users that the speed went up,the image is various and entertaining.Moreover, the proposed system can be useful for idea processing.

1．はじめに

現在，ワープロソフトを使用した文章作成が主流であり，多くの資料がワープロソフトを使用して作成されている．ワープロソフトを使用することによって文章作成は快適に行えるようになった．快適になった理由の１つめとして挙げられるのは，入力速度の向上である．キーボードで入力に慣れれば紙に記入するより格段に速度が上がる．２つめは電子的なテキストであることである．電子的なので推敲・修正が簡単，見出し・段落・重要部分などポイントをわかりやすくしやすい，テキストの再利用が自由，出来上がりをイメージしながら編集できる，などがあげられる．

2．文章作成の現状

PCを使用することによって文章入力は快適になったが，情報を探し出す作業は手間がかかる．現在，情報が必要になったときインターネット上から探してくるのが主流であるが，ウェブブラウザを開き，検索キーワードを入力して探し出すのが一般的である．この作業は一旦文章作成を中断して行うために，非常に作業効率が悪くなってしまう．しかし，上記以外の方法がないのが現状である．

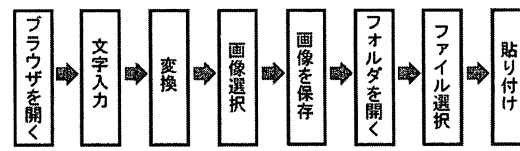

図１　画像取得の流れ

3．提案内容

そこで本論文では，日本語変換時にwebから入力文字に関連するwebコンテンツをダウンロードし，その後の変換候補選択時にダウンロードしたコンテンツを表示できるシステム「IMEdia」を提案する．

日本語変換確定時にwebコンテンツをダウンロードし，日本語変換未確定時，変換未確定文字列を元に関連したコンテンツをハードディスクから検索し表示することが特徴である．

3．1　日本語変換時にダウンロード

IMEdia をユーザは起動し，ワープロソフトなどで文章を入力する．日本語変換確定時に，形態素解析(自然言語で書かれた文を意味を持つ最小の文字列に分割し品詞に分ける作業)をして「名詞」を抽出し，抽出した名詞を別プロセスとして起動するダウンロードモジュールにコマンドライン引数として渡す．あとはダウンロードモジュールが渡された「名詞」をネット上の検索サービスでAnd 検索をし，web コンテンツをダウンロードする．ダウンロードされたwebコンテンツはフォルダごとに階層が作られそこに保存される．

3．2　未確定文字列からコンテンツを検索

ユーザが変換未確定文字列を選択中，変換未確定文字列から名詞を抽出し，ダウンロードしたwebコンテンツのフォルダを検索する．もしフォルダが見つかったなら，そのフォルダのコンテンツの内容を表示する．

3．3　コンテンツをクリップボードに転送

表示された画像などを選択し，クリップボードに転送する．画像をクリップボードに転送するには画像は一旦ビットマップ形式に変換する必要がある．が，本システムでは，選択した画像をビットマップ形式で保存し，クリップボードに転送するようになっている．

4．実装

図1，図2にシステムの流れを示す．図1はwebコンテンツの取得の流れを示し，図2は取得したwebコンテンツの表示の流れを示す．日本語入力時にIMEdiaが文字列を取得し名詞を抽出する．未確定文字列確定時はIMEdiaが取得した名詞をダウンロードモジュールに渡しダウンロードモジュールがwebコンテンツを取得する．未確定候補選択時は選択されている日本語をもとにハードディスクからコンテンツを検索し発見されれば表示する．

4．1　文字列取得

入力中の未確定文字列や確定文字列を取得するためにMicrosoft Windowsのメッセージフックの機能を用いた．また，取得した文字列から名詞を抽出するために「茶筅」を用いた．

文字列取得は IMN_CHANGECANDIDATE のメッセージをフックしている．これはIME(日本語入力システム)の変換文字列選択時に送られるメッセージであり，変換時にメッセージを取得している．また，変換確定時は，IMN_CLOSECANDIDATE のメッセージをフックすることで判断している．このメッセージは変換候補選択時に出てくるウィンドウが閉じられたときに送られるメッセージである．

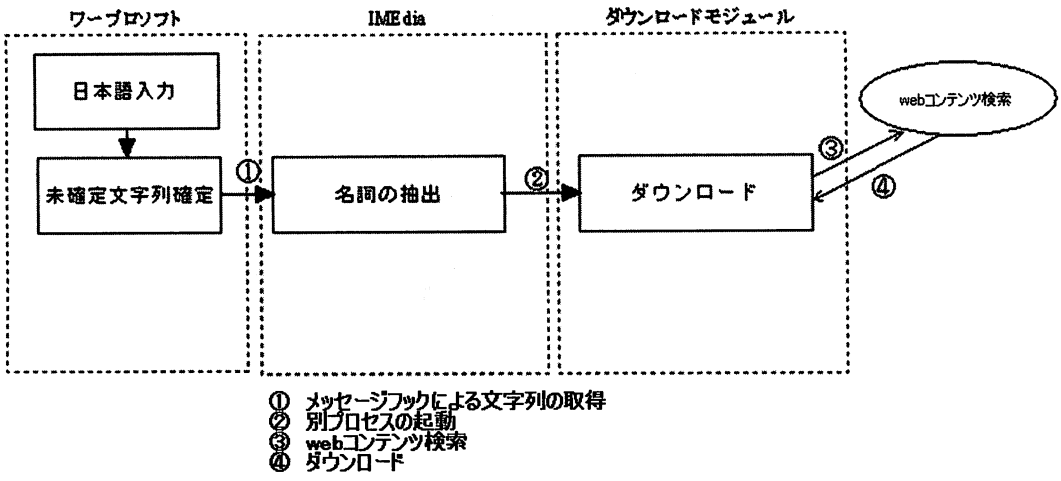

図２　画像取得の流れ

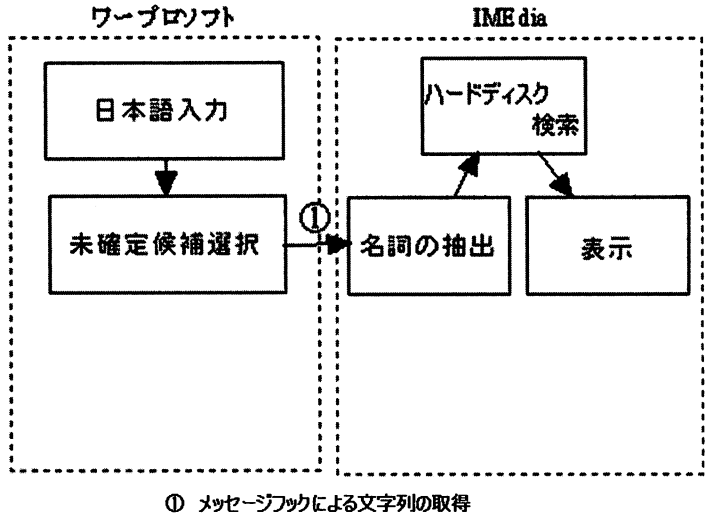

図３　画像表示の流れ

４．２　ダウンロードモジュール

Web コンテンツの取得はダウンロードモジュールによって行う．IMEdia でダウンロードを行うと，ダウンロード中は文章の入力が出来なくなり，大幅に効率が下がってしまうためである．ダウンロードモジュールを別プロセスとして起動させ，文章入力をスムーズに行えるようにした．また，ダウンロードモジュールは起動時アクティブになると文章入力の妨げになるので，起動時は非表示である．

ダウンロードモジュールは変換文字列確定時に別プロセスとして起動させる．このとき，コマンドライン引数に変換文字列から名詞を抽出したものが渡される．それらの名詞を元にインターネットサービスから web コンテンツを取得する．

４．２．１　画像取得の流れ

以下にダウンロードモジュールの画像の取得の流れを示す．

①　HTML ソースのダウンロード

ダウンロードモジュールに渡された名詞は，

google 画像検索等のインターネットサービスに And 検索がかけられる. 検索結果の HTML ソースを取得し, 正規表現により画像の URL を抽出する.

② 画像のダウンロード

ダウンロードされた HTML ソースから, 抽出された URL を元に画像をダウンロードする. 例えば「英語を映画やビジネス等」という文字列の画像をダウンロードする際は「英語」, 「映画」, 「ビジネス」, 「等」の様に名詞が分けられるので

「.¥英語¥映画¥ビジネス¥等」というディレクトリ構造でハードディスクに保存する.

また, ダウンロードされる画像は「英語, 映画, ビジネス, 等」を And 検索したものだけでなく, 「英語」, 「英語+映画」, 「英語+映画+ビジネス」, 「英語+映画+ビジネス+等」のようにそれぞれのディレクトリの階層ごとに画像をダウンロードしている.

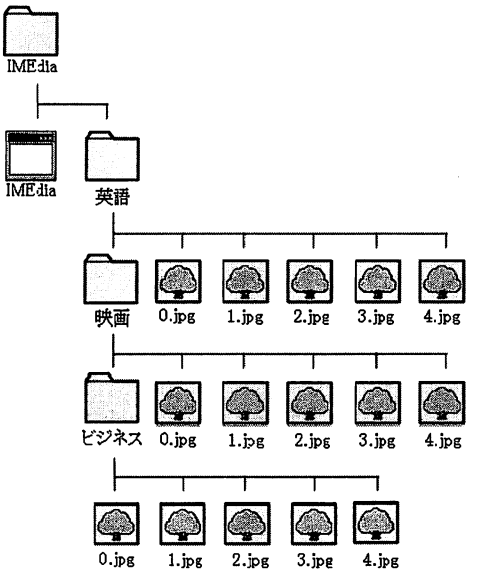

図4 フォルダの構造

③ 階層化の理由

フォルダを階層化にする理由としては図5のように画像の検索候補を絞ることが出来るからである.

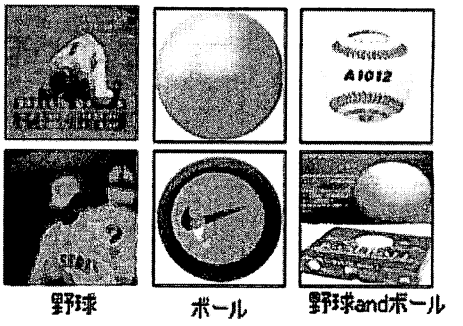

図5 画像の検索候補を絞る

4．3 取得したコンテンツの表示

取得したコンテンツは IME の変換文字列選択時に表示する. IMEdia 本体のシステムが IMN_CHANGECANDIDATE メッセージをフックしている. フックした文字列から「茶筅」を使い名詞を抽出し, ハードディスク内に検索をする. もし, コンテンツを保存したフォルダを発見出来たのならフォルダの中の画像等を IMEdia に表示させる. 図6は「野球のボール」を変換し選択した図である.

図６　IMEdia での検索候補表示

５．２　実験結果

被験者は 8 人(A～E)で行った．IMEdia を使用し文章を入力してもらい，以下の各項目に 5 段階で評価してもらった．

(5:非常に良い，4:良い，3:普通，2:悪い
1:非常に悪い)

①　使えそうな画像が出たか
②　目的の画像が出たか
③　文章にあった画像が出たか
④　画像が出ることによって文章の内容は変化したか
⑤　文章入力のスピードは通常に比べてどうだったか
⑥　表示された画像は作文時の発想支援に役立ったか

表 1 に IMEdia の使用結果を記す．

表 1　IMEdia の使用結果

	①	②	③	④	⑤	⑥
A	4	4	5	2	5	4
B	4	4	4	3	4	3
C	4	4	3	3	3	4
D	4	1	5	1	3	3
E	5	3	4	5	5	5
F	2	1	4	3	1	2
G	2	1	2	1	3	3
H	3	3	2	1	3	3
平均	3.50	2.63	3.63	2.38	3.38	3.38

①は 3.5 とある程度高い平均だったが，②は 2.63 とやや低めの平均であった．このことから画像の検索の精度を上げる必要がある．

③は 3.63 と一番高い平均であったが④は 2.38 と一番低い結果になってしまった，これにより文章にあった画像が表示されても，それが作文時の内容にあまり影響しないことがわかった．

⑤は画像のダウンロードのスピードによって変わるのであまり参考にならなかった．

⑥は画像がある程度は発想支援に役立つことがわかった．

被験者の感想

・ブラウザをわざわざひらかなくていいのでスピードはあがった.
・入力した文章によって出る画像がかわるのはおもしろかった.
・英語にも対応して欲しい
・ダウンロードに時間がかかる(初回のダウンロード)
・論文の内容を入力するのに使ったが，専門用語が多くなかなか目的の画像が出ない.

考察

実験の結果を見る限り，目的の画像はあまり出てこないが，文章にふさわしい画像はある程度出てくることが分かった．また，目的の画像が出る人と出ない人が極端に分かれることも分かった．目的の画像が出る人が必要とする画像は専門用語を含むことが少なかった．これは茶筅の辞書にはない固有名詞を判別することができないということが理由と考えられる．例えば被験者Aが「ヒヨケムシ」が出なかったとの意見が出たが，「ヒヨケムシ」を茶筅にかけてみたところ「ヒヨケムシ」は未知語として判定されていた．名詞だけではなく未知語も抽出するようにすれば固有名詞も抽出できるようになると考えられる．また,「ハイビジョンカメラ」というのは名詞を抽出できるが，これは「ハイビジョン」と「カメラ」のように 2 つの名詞が抽出されてしまう．このシステムは抽出した名詞をすべて And 検索をかけるようになっている．したがって「ハイビジョンカメラ」で検索されるのではなく「ハイビジョン+カメラ」のように検索されるので検索結果が異なる可能性がある.

６．まとめ

日本語変換確定時にコンテンツをダウンロードし，日本語変換未確定時，変換未確定文字列を元に関連したコンテンツをハードディスクから検索し表示するシステムを提案した．これにより，文章作成中にコンテンツをダウンロードし保存し，日本語変換中に表示させることが出来た．名詞の抽出には「茶筅」を用いた．提案したシステムにより，ブラウザをわざわざ開かなくても画像をダウンロードできてよいという意見や，色々な画像が取得されておもしろいというシステムの狙い通りの意見もあったが，初見では使いづらいや目的の画像があまりでてこないなどの意見もあった.

今後の課題としては

１．コンテンツを多く表示できるようにする
２．専門用語を判別できるようにする
３．コンテンツのダウンロードを早くする

などが挙げられる.

参考文献

[1] 茶筅
http://chasen.naist.jp/hiki/Chasen/
[2] Social IME
http://www.social-ime.com/
[3] Google 画像検索
http://images.google.co.jp/imghp?hl=ja&oe=UTF8-8&tab=wi&q=
[4] boost
http://www.boost.org/