データベースシステム 118－1
情 報 学 基 礎 54－1
（1999. 5. 17）

多言語知識発掘システムの構築

前田 亮，関 慶妍，植村 俊亮
{aki-mae, keien-k, uemura}@is.aist-nara.ac.jp

奈良先端科学技術大学院大学
情報科学研究科
〒630-0101 奈良県生駒市高山町8916-5

本稿では，WWWに見られるような多言語が混在する文書群を統合して管理する文書データベースの実現，さらにこれらの多言語情報への容易なアクセスを支援するシステムの実現に必要な要素技術について述べる．具体的には，単位，年号，色名など文化要素の変換を考慮した言語横断情報検索（Cross-Language Information Retrieval），文書が書かれている言語および符号系の自動識別アルゴリズム，端末に依存しないテキスト入出力について考察する．

これらの要素技術を統合することで，利用者の多言語情報へのアクセスを支援する「多言語知識発掘システム」の構築を目標としている．

Towards a Multilingual Knowledge Discovery System

Akira Maeda, Qingyan Guan, Shunsuke Uemura
{aki-mae, keien-k, uemura}@is.aist-nara.ac.jp

Graduate School of Information Science,
Nara Institute of Science and Technology (NAIST)
8916-5 Takayama, Ikoma, Nara 630-0101, Japan

This paper presents key techniques required to realize a document database which integrates a multilingual document collection typically seen in WWW, and to realize a system which supports easy access to such multilingual information. Specifically, we focus on techniques such as 1) cross-language information retrieval (CLIR) which supports conversion of cultural factors such as units, era names and color names, 2) an algorithm for automatic identification of language and coding system of documents and 3) text input/output functions which do not depend on user's terminals.

The goal of our research is to build a "multilingual knowledge discovery system" which helps users access to multilingual information by integrating these techniques.

1 はじめに

近年の世界的なインターネットおよびWWWの発展により，様々な言語で書かれた文書が提供されるようになってきている．現在のWWWで提供されている文書の約8割は英語であるが，2003年には英語以外の言語が半数以上を占めるようになると予測されている[1]．これは多言語からなる文書が混在する一つの巨大な文書データベースと考えることができる．

しかしながら，これらの多言語文書群を統一的に扱う検索システムの実現には，解決されていない様々な課題がある．例えば，問合せに利用者が用いる言語による検索対象の制限，使用される符号系の多様性の問題，端末上での問合せ文字列の入力や検索結果の表示の制限などである．

そこで本研究では，これらの課題を解決する要素技術として，

- 単位や年号など文化要素を考慮した言語横断情報検索
- 文書が書かれている言語／符号系の自動識別アルゴリズム
- 端末に依存しないテキスト入出力

について考察する．

本研究ではこれらの要素技術を統合することで，利用者の多言語情報へのアクセスを支援する「多言語知識発掘システム」の構築を目標としている．

2 関連研究

多言語情報への容易なアクセスの手段として，言語横断情報検索（Cross-Language Information Retrieval: CLIR）に関する研究が近年注目されている[1]．これは，ある言語で書かれた文書群を別の言語による問合せで検索するものである．この技術は，検索対象の文書群自体が多言語である場合や，利用者が検索対象となる文書の言語を読むことはできるが，問合せを記述するのが困難である場合などに有効である．

言語横断検索に用いられる手法は大きく分けて，検索対象の文書群を翻訳する方式と，問合せを翻訳する方式の二つがある．前者は，既存の機械翻訳システムを用いることができ，文脈を考慮できることにより訳語の曖昧性も低くなることから一般に後者より高い検索性能が得られるが，WWWのような大規模な文書群に対しては現実的ではない．後者の場合は，一般に問合せは短く単に単語の羅列からなる場合もあるため，訳語の曖昧性の除去が問題になるが，既存の検索エンジンをそのまま使えることから，主にこの方式が研究の対象となっている．問合せ翻訳方式の言語横断情報検索へのアプローチとしては，対訳辞書，多言語シソーラス，あるいは並列コーパスを用いる手法などがこれまでに研究されている[2, 3]．

一方，文書の言語／符号系の識別に関しては，ヨーロッパの諸言語についてn-gramや単語辞書などを用いる言語識別の手法が研究されている[4]．アジア圏の言語については，コーパスによる統計を用いて15言語11符号系を自動識別するアルゴリズムが提案されており，約95%の正解率が得られている[5]．ただしこのアルゴリズムでは，一つの文書に対して対応するすべての言語／符号系を仮定して計算する必要があるため，対応言語／符号系が多くなると効率が落ちる．

3 システム構成

本システム全体の構成を**図1**に示す．

本システムは，**文書収集部**，**索引付け部**，**検索部**からなっている．

文書収集部では，Webロボットを用いて文書を収集する．Webロボットとは，リンクを再帰的に辿ることによってWWWのハイパテキスト構造を自動的に取得するソフトウェアである．

索引付け部では，収集した文書を後述の言語／符号系識別モジュールにより識別し，必要に応じて符号系の変換を行い，対応した言語別の索引付けモジュールで索引付けを行い，索引データベースに格納する．

検索部では，利用者からの問合せを翻訳エンジンにより利用者が指定した各言語に翻訳し，翻訳された問合せを用いて検索エンジンにより検索を行い，最終的に検索結果が利用者に返される．

[1] Nua Internet Surveys, `http://www.nua.net/surveys/`

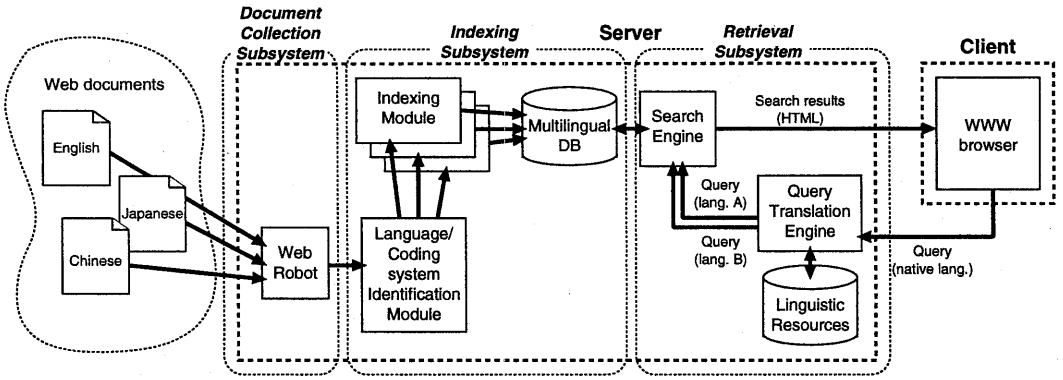

図 1： 本システム全体の構成

4 多言語検索システムの要素技術

4.1 言語横断検索

4.1.1 問合せの翻訳

既存の言語横断検索手法では，対訳辞書に加えて，訳語の曖昧性の除去のために並列コーパスやコンパラブルコーパス[2]などが用いられるが，これらは容易には入手できない．また，対訳辞書についても，言語対によっては電子化されたものが入手できない場合も考えられる．本研究では，入手できる言語資源にできるだけ依存しない方法を考える．直接目的言語に対する対訳辞書がない場合は，第3の仲介言語（例えば英語）を介した翻訳も考慮する（図2）．

この場合，当然検索性能の著しい低下は避けられないが，このような検索に対するニーズは少なからず存在すると考える．

本システムでは，基本的に原言語（問合せの言語）から目的言語（検索対象の言語）への対訳辞書を翻訳に用いる．直接原言語から目的言語への対訳辞書がない場合は，前述のように仲介言語を介した翻訳を行う．

[2]文ごとあるいは記事ごとに対応づけられた二言語のコーパス

4.1.2 曖昧性の除去

対訳辞書により得られた訳語の候補は，目的言語のコーパスにおける単語間の共起頻度を用いた以下のような単純な手法で曖昧性の除去を行う．

原言語による問合せの単語集合を

$$S = s_1, s_2, \ldots, s_m$$

原言語による問合せの単語 s_pに対する目的言語における訳語の候補集合を

$$T_p = t_{p1}, t_{p2}, \ldots, t_{pn_p}$$

とする．ここで n_pは問合せの単語 s_pに対する訳語候補数である．

$f(w_1, w_2)$ を単語 w_1と w_2の目的言語のコーパスにおける共起頻度とすると，訳語として選択される単語 t_pは，

$$t_p = \max_{q,r} f(t_{pq}, t_{mr}) \quad (q = 1 \ldots n_p, r = 1 \ldots n_m)$$

により得られる．

コーパスには，文書収集部によって収集され言語識別モジュールによって言語別に振り分けられた文書群を用いることを考えている．

また，単語の品詞情報を用いることが翻訳の曖昧性の除去に有効であると考えられることから，上記手法との併用を検討している．

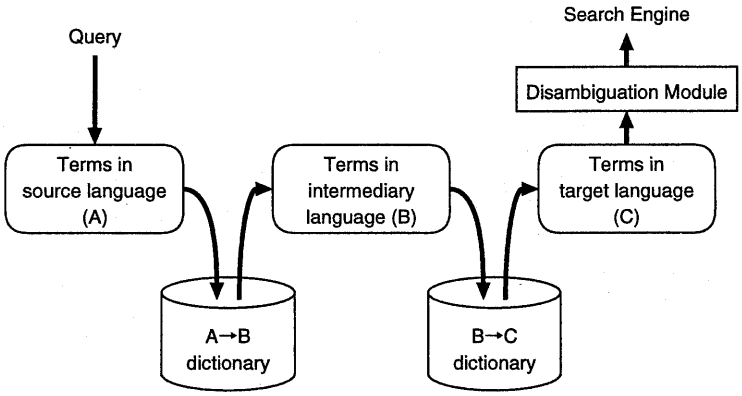

図 2： 仲介言語の辞書を介した翻訳

4.1.3 文化要素の変換

既存の言語横断検索手法では，基本的に問合せには一般的な単語あるいはフレーズを想定しており，問合せに現れる単位や年号など文化要素の変換までは考慮されていない．また，色の名前など言語や文化固有の表現は，容易に目的言語に翻訳することができない場合がある．例えば，「平成 10 年に発売された 10 万円以下のパソコン」という問合せの翻訳は，従来の言語横断検索手法では対応できない．このような変換は，問合せの同義語展開の言語横断検索への拡張と考えることができる．

ただし現状の HTML 文書は構造化が不十分であり，文書中に現れる日付や金額などはすべて単なる文字列としてしか扱えない．しかしながら，将来 HTML に代わって構造化が強化された XML (eXtensible Markup Language) [6] が普及することで，上記の例のような問合せも重要になってくると思われる．

これら文化要素の変換の実装を考えると，度量単位や年号，色の名前などは静的であるため数式や変換表によって容易に変換が可能であるが，通貨単位は動的であるため，そのような方式では対応できない．この場合，外部の WWW サイトなどから為替レートなどの動的な変換表を得ることが考えられる．また，上記の例のように過去の通貨単位の変換が必要になる場合も考えられる．この場合も，過去の任意の日付の為替レートを提供する WWW サイトが存在すれば，これを用いることができる．

4.2 言語/符号系の識別

4.2.1 WWW における文書の符号系の問題

WWW で提供される文書は様々な言語や符号系が用いられている．インターネットで使用される符号系を登録する機関である IANA (Internet Assigned Numbers Authority) には 214 もの符号系が登録されている[3]．

HTML 文書に符号系の情報を付与する方法として，HTTP の Content-Type ヘッダ中の charset パラメータで指定する方法，同様の指定を HTML の META タグを用いて文書内で行う方法 [7] があるが，現実には符号系についての情報は付与されていない場合が多い．

また，Unicode[8] のような多言語対応の符号系が今後普及したとしても，その中での言語の識別は依然必要である．このため，これらの自動識別技術は，WWW 情報の検索システムに必須の技術として位置付けられる．本システムでは，言語に応じた索引付けを行うためにこの技術を用いる．

我々は，より単純で効率的な識別アルゴリズムとして，符号系に関わらず 1 バイトのコード分布の統計を用いる方法を検討している [9].

[3] http://www.isi.edu/in-notes/iana/assignments/character-sets

表 1： 言語／符号系識別の対象とする符号系

符号系	言語	単位	図形文字領域	ピーク
ASCII	英語	7	33-126	-
ISO-2022 (7bit)	多言語	7	33-126	-
Shift_JIS	日本語	8	33–252	129–131
EUC-JP	日本語	8	33–126, 142–254	161–165
GB2312	中国語（簡体字）	8	33–126, 161–254	161–254
Big5	中国語（繁体字）	8	33–126, 161–254	64–126, 161–180
EUC-KR	韓国語	8	33–126, 142–254	161–254

4.2.2 対象とする符号系

提案する識別アルゴリズムは，現時点では**表 1** に示す 4 言語 7 符号系を対象としている．

表中の**単位**は 1 バイト中で実際に使用されるビット数を，**図形文字領域**は図形文字が割り当てられているコード領域を，**ピーク**は後述の統計データにおいて出現頻度が比較的高かった領域を表している．

上記のうち ASCII 以外の符号系はすべて多バイトコードの符号系であるが，本アルゴリズムでは符号系に関わらず，すべて 1 バイトコードとして考える．

4.2.3 統計データ

コード分布を分析するための統計データとして，各符号系ごとに約 100K バイト分の文書を WWW から収集し，これを用いた．

ただし，ASCII と ISO-2022 については，7 単位系であることから最上位ビットで他の 8 単位符号系との判別ができ，二つの区別はエスケープシーケンス（コード値 27）の有無により判定できるため，統計は取っていない．

各符号系における 1 バイトコードの出現頻度の分布を**図 3** に示す．

4.2.4 自動識別アルゴリズム

本アルゴリズムでは，まず対象となる文書の 1 バイト毎のコード値（0〜255）の出現頻度を計算する．次に，統計データにおける頻度分布の特徴からヒューリスティックに作成した識別アルゴリズムを適用する．アルゴリズムを **Algorithm 1** に示す．

ここで，入力は識別する文書の頻度分布であり，出力は識別された符号系である．$freq(x)$ はコード値 x の出現頻度であり，$avg\_freq(m \ldots n)$ はコード値の範囲 m から n 内の出現頻度の平均を表すものとする．

if $avg\_freq(33 \ldots 126) = 0$ **then**
  if $freq(27) > 0$ **then**
    return “*ISO-2022*”
  else
    return “*ASCII*”
  end if
else if $avg\_freq(128 \ldots 141, 143 \ldots 160) \neq 0$ **then**
  return “*Shift_JIS*”
else if $\frac{avg\_freq(166 \ldots 180)}{avg\_freq(181 \ldots 254)} > 2.0$ **then**
  return “*Big5*”
else if $\max(freq(164), freq(165)) > 0.02$ **then**
  return “*EUC-JP*”
else if $freq(192) > 0.02$ and
$avg\_freq(161 \ldots 170) > 0.02$ **then**
  return “*EUC-KR*”
else
  return “*GB2312*”
end if

Algorithm 1: 符号系の自動識別アルゴリズム

4.2.5 実験結果

WWW から無作為に収集した，各符号系につき 100〜251 件，合計 1389 件の文書（1 文書 36 バイト〜221K バイト）について上記アルゴリズムによる識

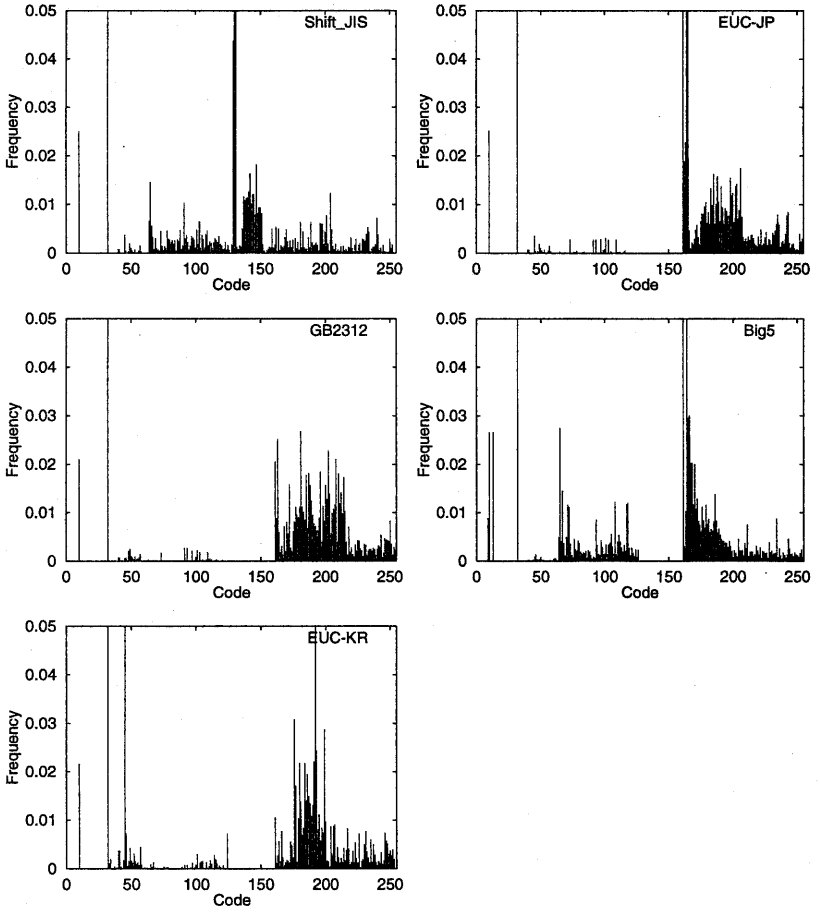

図 3： 各符号系における 1 バイトコードの分布

別結果の正解率を調べる実験を行った．実験環境は Sun Ultra-1（UltraSPARC 167MHz, Solaris2.5.1）を用い，全文書の処理に 24.8 秒，一文書平均で約 0.02 秒を要した．この実験の結果得られた正解率を**表 2** に示す．

誤って識別された文書は合計で 14 件あった．これらは，

- 文書（HTML タグ以外の部分）が短い
- ASCII 以外の文字（2 バイト文字）が極端に少ない
- ひらがなの出現頻度が少ない（人名や組織名のリストなど）

などの場合であった．

表 2： 符号系識別アルゴリズムの正解率

符号系	正解率
ASCII	100.0%
ISO-2022	100.0%
Shift_JIS	100.0%
EUC-JP	96.4%
GB2312	98.0%
Big5	99.5%
EUC-KR	99.0%
平均	99.0%

4.3 索引付け

文書の索引付けは，例えばヨーロッパ言語の場合は stemming や不用語の除去 [10]，日本語や中国語などでは単語への分割 [11] など，言語に依存した処理が必要になる [4].

本研究では，一文書中に複数の言語が混在する場合についても考慮するため，各言語による索引を統合的に管理する手法を検討している．

本システムにおける索引付け部分の構成を図4に示す．WWW から収集した文書は，まず言語／符号系識別モジュールで各言語に振り分けられ，必要であれば符号系の変換が行われる．次に識別された言語によって，対応する索引付けモジュールに渡され，言語依存の索引付けが行われる．最後にその結果が索引データベースに格納される．

既存の索引付け手法は一言語のみを対象としているが，Unicode などの一文書中に複数の言語が混在する文書の適切な索引付け手法についても今後検討する必要がある．

4.4 テキスト入出力

利用者が多言語情報にアクセスする際の別の問題として，テキスト入出力の制限が挙げられる．例えば，日本語のフォントや入力メソッドは基本的に日本語版の OS にしかインストールされていないため，他の国のコンピュータから日本語を表示／入力するためにはそれらをインストールする必要がある．

この問題を解決する手段として，図書館情報大学で開発された多言語 HTML ブラウジングシステム（MHTML）[12] がある．このシステムでは，表示／入力に必要なフォントを HTML 文書に付加し，クライアントでの表示／入力には Java アプレットを用いることで，利用者によるフォントなどのインストールを一切必要とせずに様々な言語の表示／入力が可能となる．

本システムでは，これをテキスト入出力のインタフェースとして用いることで，多言語情報へのより容易なアクセスを実現する．

MHTML を用いた利用者インタフェース部分の構成を図5に示す．

5 おわりに

本稿では，多言語文書群を統一的に扱う検索システムの実現に必要な要素技術のいくつかについて考察し，これらを統合する「多言語知識発掘システム」の構想を述べた．

本研究の今後の課題としては，

- 言語横断検索手法の実装および検索性能の評価実験
- 言語／符号系の自動識別技術における対応言語／符号系の拡大
- 多言語混在文書の索引付け手法の検討
- 検索結果の提示手法の検討
- システムの実装および評価

などが挙げられる．

参考文献

[1] Klavans, J. L. and Schäuble, P.: NSF–EU Multilingual Information Access, *Communications of the ACM*, Vol. 41, No. 4, p. 69 (1998).

[2] Oard, D. W.: Alternative Approaches for Cross-Language Text Retrieval, *Electronic Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval* (1997).

[3] Grefenstette, G.(ed.): *Cross-Language Information Retrieval*, The Kluwer International Series on Information Retrieval, Vol. 2, Kluwer Academic Publishers (1998).

[4] Wechsler, M., Sheridan, P. and Schäuble, P.: Multi-Language Text Indexing for Internet Retrieval, *5th RIAO Conference, Computer-Assisted Information Searching on Internet (RIAO'97)* (1997).

[5] Kikui, G.: Identifying the Coding System and Language of On-line Documents Using Statistical Language Models, 情報処理学会論文誌, Vol. 38, No. 12, pp. 2440–2448 (1997).

[6] World Wide Web Consortium: Extensible Markup Language (XML) 1.0 (1998). W3C Recommendation 10-February-1998
http://www.w3.org/TR/1998/REC-xml-19980210.

[7] Yergeau, F., Nicol, G., Adams, G. and Düerst, M.: Internationalization of the Hypertext Markup Language, RFC 2070, Network Working Group (1997).

[8] Unicode Consortium, T.: *The Unicode Standard, Version 2.0*, Addison-Wesley (1996).

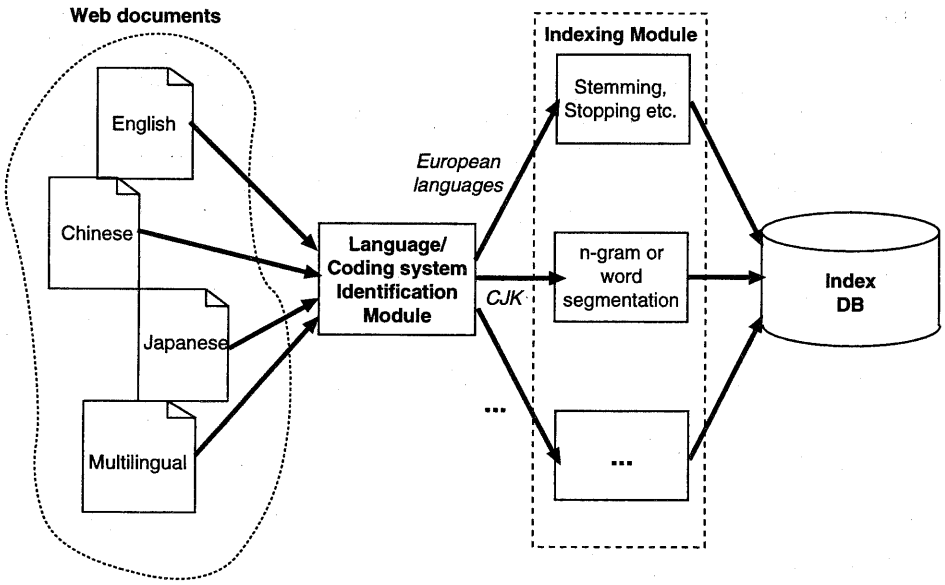

図 4： 索引付けの流れ

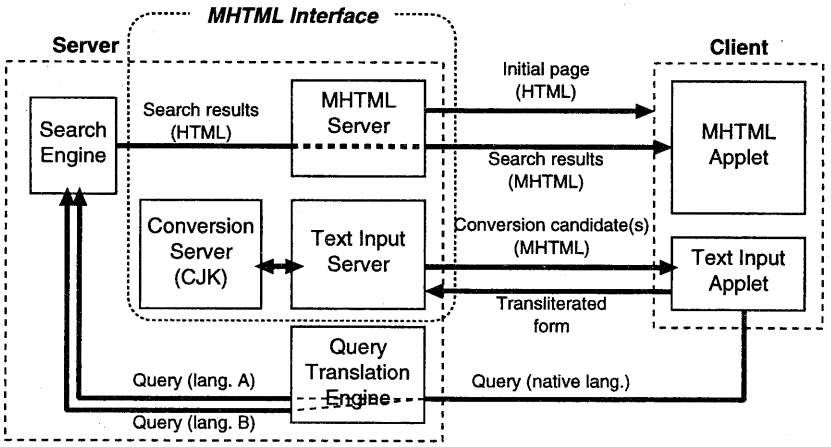

図 5： 利用者インタフェース部分の構成

[9] 前田亮, 関慶妍, 植村俊亮: 1バイトコード分布に基づく文書の符号系の自動識別, 1999 年電子情報通信学会総合大会講演論文集, D-5-3, p. 77 (1999).

[10] Frakes, W. B. and Baeza-Yates, R.(eds.): *Information Retrieval – Data Structures & Algorithms* –, Prentice-Hall (1992).

[11] Fujii, H. and Croft, W. B.: A Comparison of Indexing Techniques for Japanese Text Retrieval, *Proceedings of the 16th Annual International ACM SIGIR Conference (SIGIR'93)*, pp. 237–246 (1993).

[12] 前田亮, Dartois, M., 太田純, 藤田岳久, 阪口哲男, 杉本重雄, 田畑孝一: クライアントにフォントを必要としない多言語 HTML 文書ブラウジングシステム, 情報処理学会論文誌, Vol. 39, No. 3, pp. 802–809 (1998).