XMLデータベースからの動的Webページ生成環境における変更検出・通知方式

神戸大学
宮崎慎也　　馬 強
京都大学
田中克己

研究背景

- Web上の情報量の増大

  頻繁なページ更新，追加

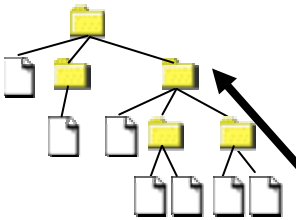
ユーザ：常に重要な情報を獲得することが
困難な場合がある
サーバー：新規情報をユーザにアピールしたい

変更通知機構が有用

今までの研究:WebSCAN
(Web Site Change Analyzer and Notifier)

Webの大量な変更から重要な情報をPick Upして，通知(Pus型)するシステム([宮崎2001], TOD10)

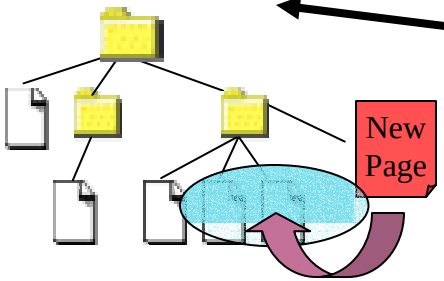
www.nikkansports.co.jp



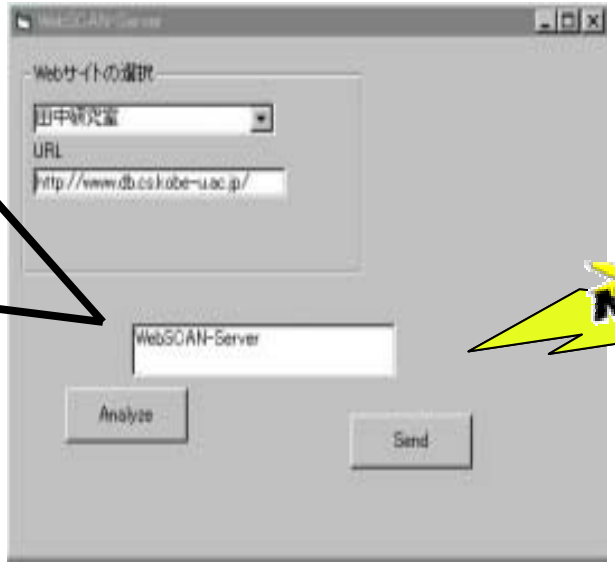
監視

- 変更解析
- 変更情報作成，配信

www.db.cs.kobe-u.ac.jp



時間および内容の比較
(新鮮度･流行度)



Server

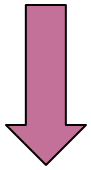
変更通知

Client

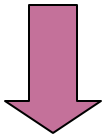
Personalizedユーザビュー(XML/XSL)

WebSCANの課題

動的Web**ページが多数存在**（CGI,ASP,JSP,etc.），One-to-One**の**Web**サイト**

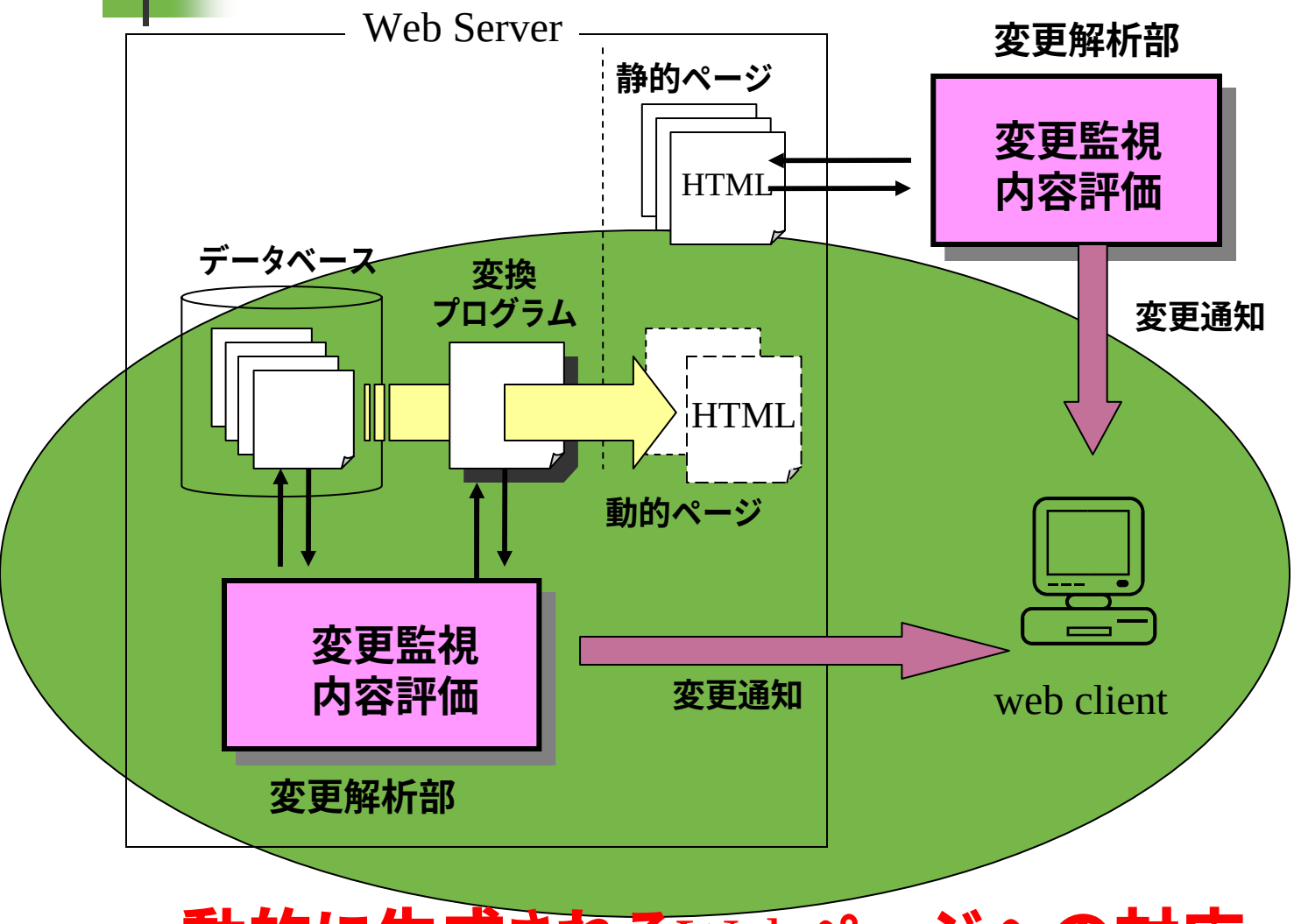


- **動的ページ未対応**
- **サーバー側技術への対応**



サーバー内で動的ページの変更監視･評価および通知

システムモデル(WebSCAN-Ⅱ)



動的に生成されるWebページへの対応

アプローチ

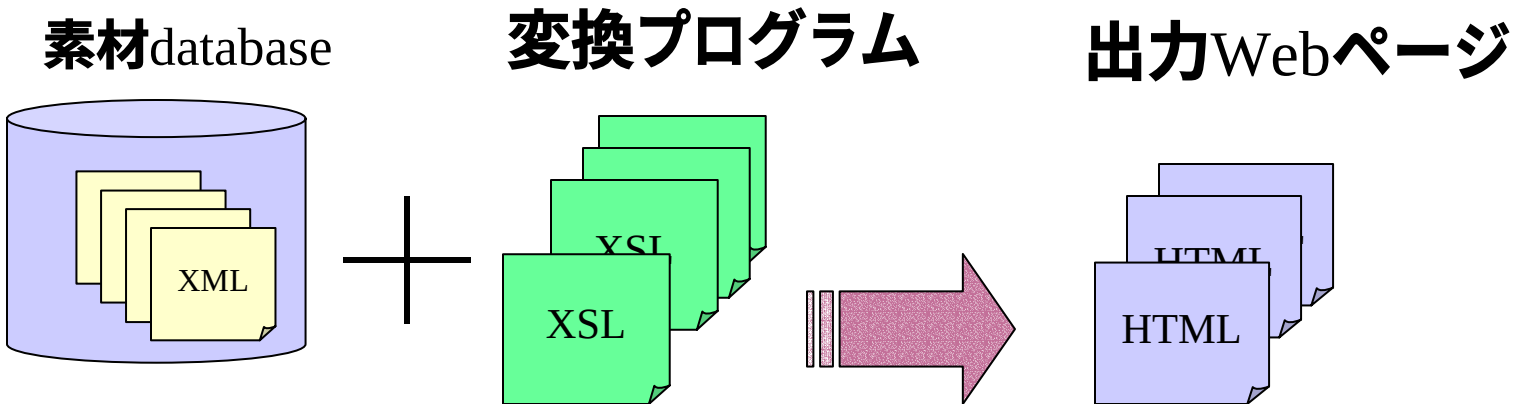
- 素材データ (XML DB) と変換プログラムの両方を監視・解析

  → 変更の検出

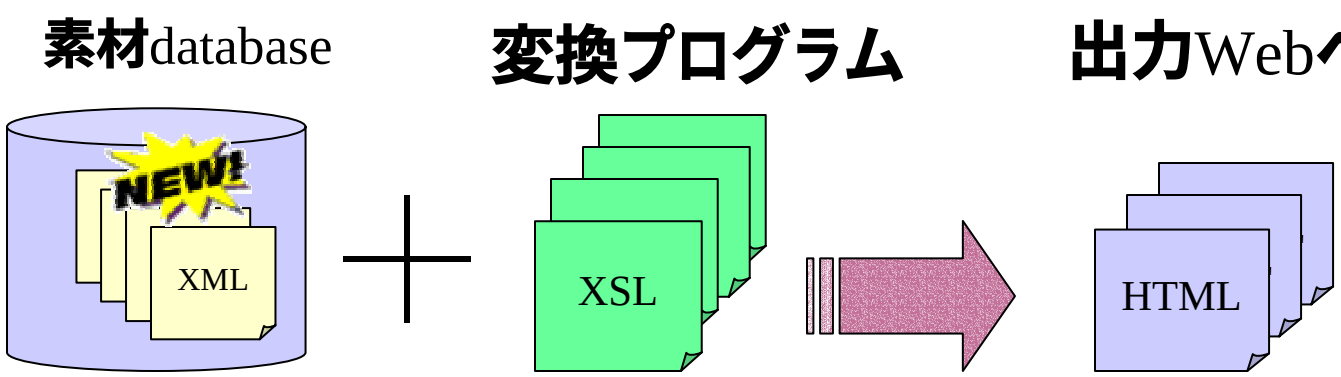
- 内容，構造とスタイルを考慮して変更評価・通知
  - XMLデータの構造差異
  - 内容差異
  - スタイル・レイアウト差異
- PUSHペースの変更通知
  - クライアント側で変更通知の個別化

評価する変更の検出

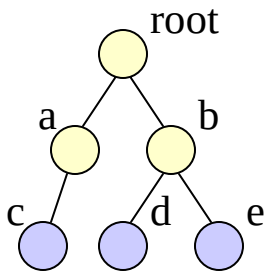
- **監視すべき変更パターンの検出**
  変換プログラムの解析より出力に影響を与える変更パターン(CP)を抽出
- **実際の変更**
  素材データと変換プログラムの変更(CS)
- **評価する変更(EC)**
  EC = (cs|cs ∈ CS, pattern(cs) ∈ CP)



XMLデータの解析(1/2)



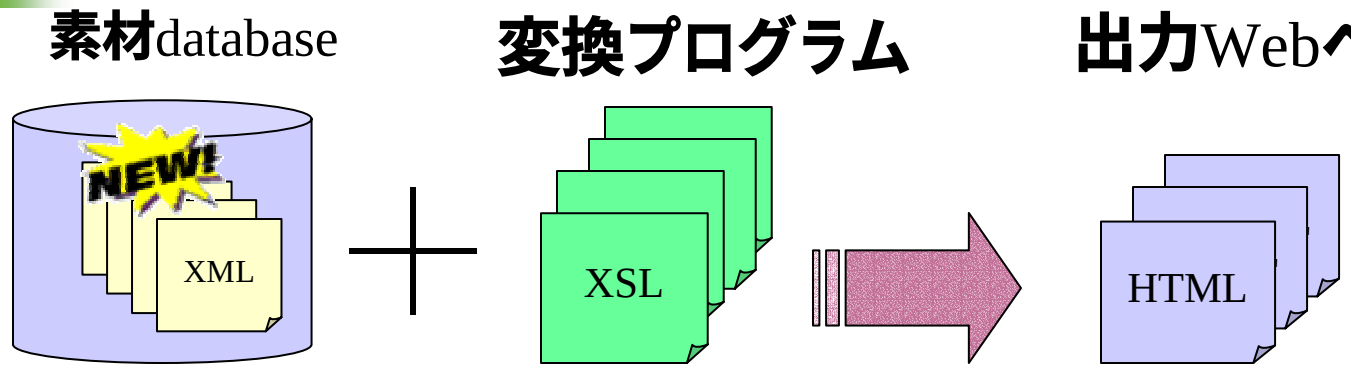
- 内容の評価



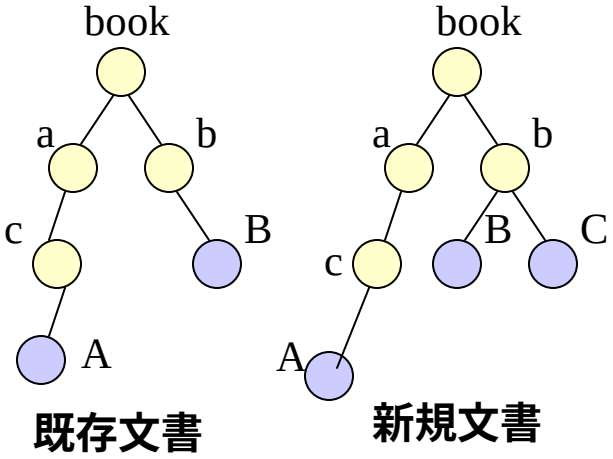
- テキストノード：
  新鮮度，流行度を評価
- 数値ノード：
  偏差値を用いて評価

比較対象は過去の文書中の同一ノード

XMLデータの解析(2/2)



•構造の差異検出



追加文書中の各テキストノードに対し，
XPathを用いて，過去の文書へ問い合わせ

- /book/a/c/A
- /*/book//a//c//A
- /*/book/A,/*/a/A, /*/c/A,/*/a/c/A

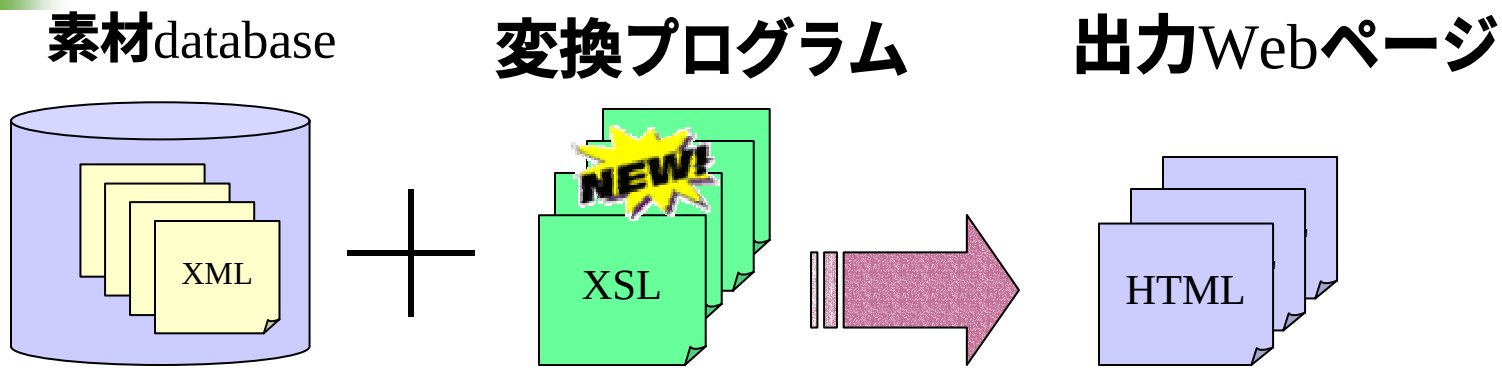
ノードCの特徴量 S (c) :

$$S(c) = 1 - \frac{m_c}{n}$$

m_c ← 問い合わせ結果数 (類似しているノード数)

n ← 総文書数

変換プログラムの解析



- 監視する変更パターンの更新
  - ページに出力されるデータ（ノード）を特定
- ページスタイルの変更を変更通知に利用
  - ⟶ ページのスタイル変更を視覚的に通知

変更通知(配信コンテンツ)

- XMLで記述
  - 変更評価値
    - 新鮮度，流行度
    - 変更ノートの特徴量
  - 変更内容
    - 最初のテキストデータ(ノード)
    - 最初のイメージファイル(ノード)
    - URL

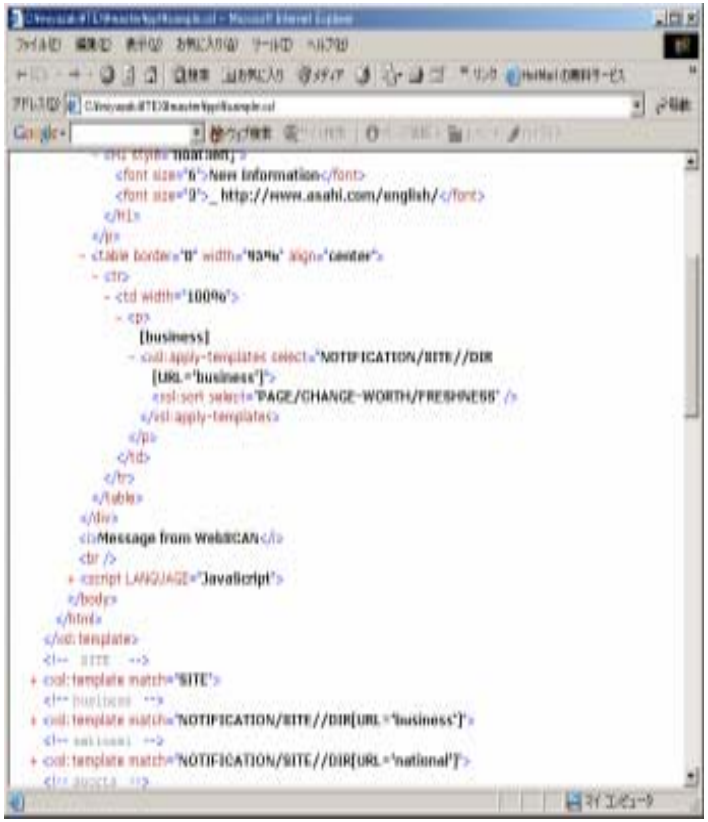
(記述例)

```
<?xml version="1.0" encoding="UTF-8" ?>
- <NOTIFICATION>
  - <SITE>
      <URL>http://www.asahi.com/english/</URL>
    - <DIR>
        <URL>business</URL>
      - <PAGE>
          <URL>K20010822005 78.html</URL>
        - <CHANGE-WORTH>
            <FRESHNESS>0.687</FRESHNESS>
            <POPULARITY>0.354</POPULARITY>
            <UPD-FREQ>1</UPD-FREQ>
          </CHANGE-WORTH>
          <TITLE>asahi.com : ENGLISH</TITLE>
          <INDEX>Rightist publisher preys on discrimination
            fears</INDEX>
          <A>Rightist publisher preys on discrimination
            fears</A>
          <A>Job cuts are no sure bet in ailing electronics
            sector</A>
          <A>Ministry eyes property tax cuts</A>
          <A>Japan-China trade hits record high in Jan.-
            June period</A>
          <A>Brazil's phone war yields ultra-low rates</A>
          <ABST>Even the Defense Agency bought
            expensive books to avoid being branded
            insensitive. A publishing house whose de facto
            owner is a leader of a right-wing group raked in
```

変更通知の個別化(1/2)

- ユーザプロファイルを XSLで記述

(記述例)



- 通知情報のフィルタリング
- 変更呈示のレイアウト
  (変更がスタイル変更の場合，変更したスタイルを利用)
- 最終的な価値計算

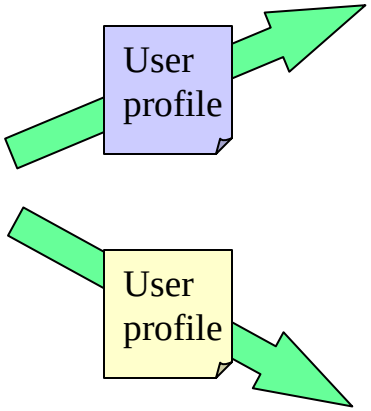
↓

XSLTによって個別化された通知に変換

変更通知の個別化(2/2)

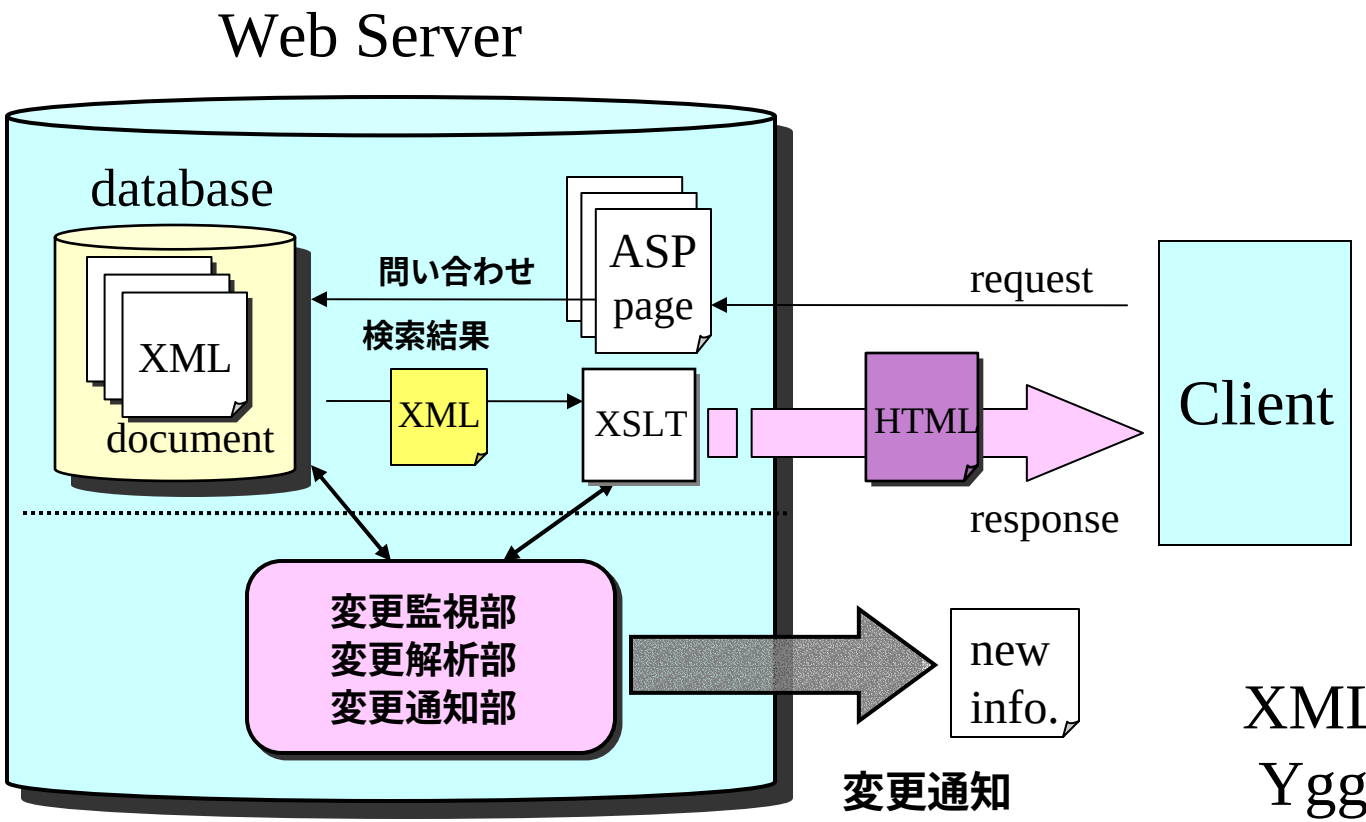
- 個別化された変更情報の呈示

同一の変更（配信）情報



ユーザごとの変更呈示

プロトタイプシステム



XML**サーバー**：
Yggdrasill 1.5
Web**サーバー**：
IIS 5.0

おわりに

- Webページ，特に動的Webページの変更通知機構を提案
  - ソースデータと変換プログラムの両方監視･解析
  - 内容･構造･スタイルを考慮した変更評価･通知
  - 個別化可能な変更通知機構
- 今後の課題
  - XMLデータの構造差異の解析･評価
  - 変換プログラムによる出力ページの構造変化に対する解析
  - 応用アプリケーション（EC分野，etc.）