2J-1

# 自然言語による e-learning 用 Web ページ検索システム

†澤井 進　　‡森 勇喜　‡若木 利子　‡板倉 弘幸
(財)学習ソフトウェア情報研究センター　芝浦工業大学

## 1. はじめに

現在，群馬県生涯学習情報提供システム「まなびねっとぐんま」[1]では同県内で開催されている「おもしろ科学教室」について記述されている。**(財)**学習ソフトウェア情報研究センター[2]ではこの科学教室をデジタル撮影し, e-learning の一環として Web 上で公開を行うプロジェクトが進行中である．

本研究では，上記プロジェクトの一つとしておもしろ科学教室の情報を掲載した Web ページ・データベースを自然言語入力で検索する Web アプリケーションの実現を行う．Web ページのテキストシナリオ部分からデータベースを作成し，潜在的意味インデキシング（Latent Semantic Indexing：LSI）の検索手法[3, 6]を使い, 入力質問文章と類似の Web ページを検索する．テキストシナリオには該当ページに掲載されている動画や静止画からなるデジタル映像の内容が説明されており，これを利用することで質問に類似した映像の検索を, テキスト検索で実現する．

また，動画や静止画による暗黙知は文章による形式知よりも直感的に人間の記憶に植えつける手段としてその能力を発揮し，自分や他人のコンテキスト（状態）を理解する手段として最適である．このことから質問文章と検索結果の暗黙知を利用することによって、容易に得られる類似性の良し悪し判断をユーザが行う。その情報をシステムに与えて検索結果を改善する適合性フィードバックが有効と考えられるので, これを実装したシステムの実現を目指す．

## 2. 本システムの特徴

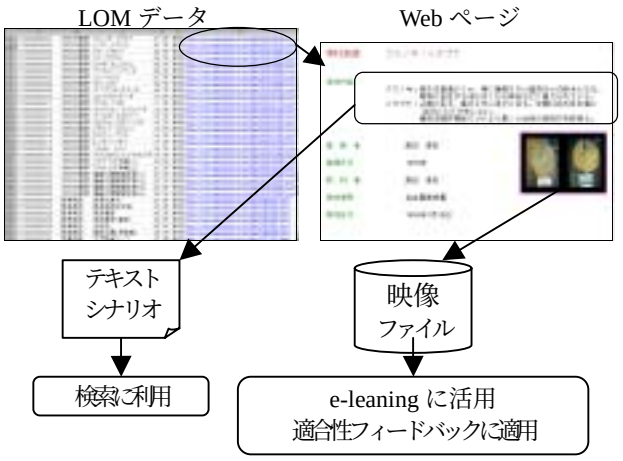


図１．システムの特徴

検索にはインターフェイスとしてデジタルアーカイブの検索を可能にするメタデータ情報（Learning Object Metadata: LOM）を使用する．これを利用することによって NICER（National Information Center for Educational Resources：教育情報ナショナルセンター）[4]の検索システムへの利用が可能となる．

検索対象となる Web ページは自然言語であるテキストシナリオと動画や静止画のデジタル映像ファイルを持つ．このテキストシナリオは映像ファイルの内容が記されており，これをデータベース化することで検索を可能とする．また，映像ファイルにおいては暗黙知を利用して e-learning に活用するとともに，フィードバック検索を行う一助となる．

検索に用いる質問は自然言語による文章入力とする．これにより，一般性が高くなり，多くの利用価値を得られる．

## 3. 本システムの概要

本検索システムは構成図を以下に示す．

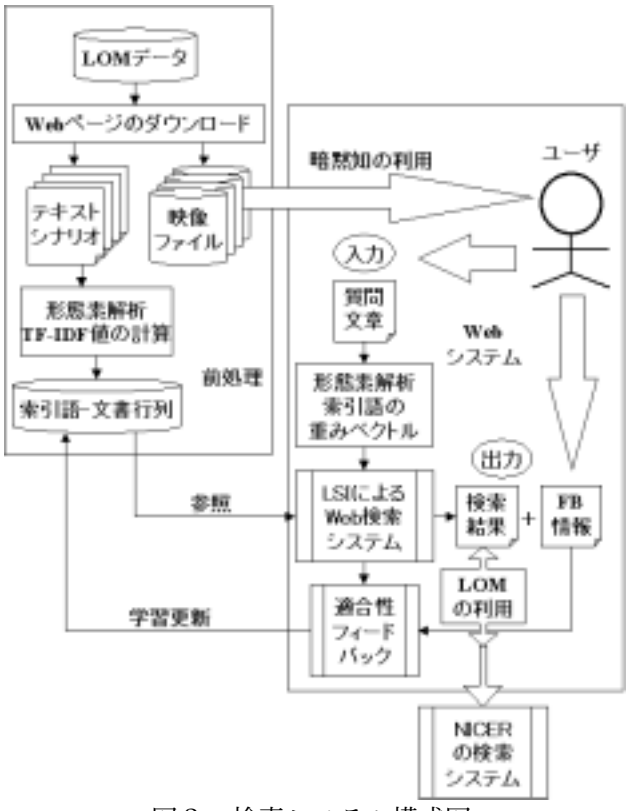


図２．検索システム構成図

### 3.1 索引語-文書行列の構築

LOM データから ID ナンバー，タイトル，Web ページの URL などの必要な情報を取り出しファイルに格納するとともに, その URL からテキストシナリ

Web-based e-learning search system using natural language query
† Susumu Sawai
Software information research center for Learning
‡ Yuhki Mori,Toshiko Wakaki,Hiroyuki Itakura
Shibaura Institute of Technology

オをダウンロードし，形態素解析，TF-IDF 値の計算により索引語を抽出した後, 索引語-文書行列を構成する．

・形態素解析

名詞が Web ページの特徴を表すものと想定し，テキストに対して日本語形態素解析ツールである茶筌[5]を用いて形態素解析を行い，名詞だけを取り出す．数字や代名詞は stoplist として削除する．

・TF-IDF 値の計算

各ページで名詞がどのくらいの頻度で出現したかを表す値（TF）と単語が全文書中のどれくらいの文書に出現するかを表す値（IDF）により，単語の重み値（TF-IDF）を導き，その値がある閾値以上の単語を索引語として抽出する．

$$tfidf(d,t) = tf(d,t)\cdot idf(t) = tf(d,t)\cdot \log\frac{N}{df(t)} + 1$$

$N$ : 全文書数、$df(t)$ : $t$が出現する文書数

・索引語-文書行列の構築各 Web ページは，抽出された索引語を次元としたベクトル空間における文書ベクトルで表現され, その要素が TF-IDF 値である．これらの文書ベクトルの集合を行列表現したものが索引語-文書行列である．

### 3.2 LSI 法による Web 検索システム

自然言語で表現された質問文書に対して索引語の重みベクトルを生成し，LSI 法により索引語-文書行列を参照して類似度の高い Web ページを検索・出力する．

・潜在的意味インデキシング(LSI)

潜在的意味インデキシングは索引語-文書行列に対して特異値分解を行い, 特異値の小さい次元を圧縮して元の索引語-文書行列より低いランクの基底行列を求める．その基底に各ベクトルを射影することにより，文書ベクトルの次元を圧縮する方法である．

索引語-文書行列を $m\times n$ 行列 $D$ とする．$D$ の特異値分解は次のように定義することができる．

$$D = U\Sigma V^T$$

$U$ は $m\times m$ 直行行列($UU^T = U^TU = I$)，$V$ は $n\times n$ 直行行列($VV^T = V^TV = I$)，$\Sigma$ は $m\times n$ 行列である．$rank(D) = r$ とすると, 対角線上に $r$ 個の特異値 $\sigma_i$ が大きさの順に並んだものである．

$r$ 次元の文書ベクトル $d$ を $k$ 次元の文書ベクトル $d^{(k)}$ で近似するためには，$U_k$ の張る空間への射影を考えればよい．$U_k$ は $U$ の最初の $k$ 個($k < r$)の左特異ベクトルのみの $m\times k$ 行列である．

$$d^{(k)} = U_k^T d$$

このようにして，索引語-文書行列を低次元に圧縮する．

ベクトル空間での文書ベクトル $d$ と質問ベクトル $q$ に対して，余弦 $sim(d,q)$ を取ることにより，類似度を求める．類似度 $sim(d,q)$ は k 次元に圧縮した索引語-文書行列 $D_k$ を求めずに $U_k$，$\Sigma_k$，$V_k$ から計算できる．

### 3.3 適合性フィードバックシステム

検索された結果の精度の良し悪しをユーザが判断し，その情報をシステムに与えた適合性フィードバックよって，検索結果の改善を行う．

## 4. 検索システムの実現

本システムは現在，芝浦工業大学若木研究室のサーバにて稼動中である。質問文章を入力することにより容易に Web ページ検索ができる．（図３）

C 言語で実現し，特異値，固有値の計算には CLAPACK を用いた. Web では C 言語による CGI を用いて実現した．適合性フィードバック以外は，ほぼ実装済みである．

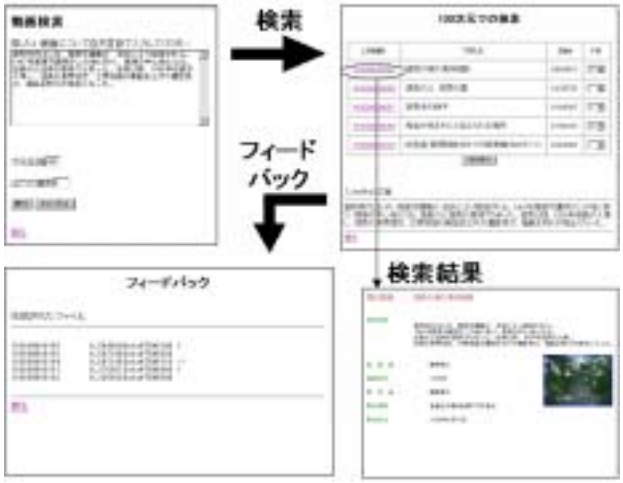


図３．Web での検索イメージ

## 5. おわりに

検索結果においてはまずまずの結果を得ることが出来た．しかし，現在はデータ量が２００弱と少ないので，今後はデータ量を増やし，検索精度の評価を検討したいと思う．また，適合性フィードバックシステムの Web 上の導入も実現していく．

## 参考文献

[1] まなびねっとぐんま
http://www.manabi.pref.gunma.jp/
[2] 学術ソフトウェア研究センター
http://www.gakujoken.or.jp/
[3] 北 研二,津田　和彦,獅々堀　正幹：情報検索アルゴリズム，共立出版，pp.51-89，2002
[4] http://www.nicer.go.jp/
[5] 茶筌：http://chasen.aist-nara.ac.jp/
[6] 督永，樋口，若木，新田：判例データベースからの類似文書検索システムの開発と評価，平成 15 年度電子情報通信学会東京支部学生会研究発表会講演論文集，D-8，p.108，2004．