LSI + k-means クラスタリングを用いたデジタルアーカイブ検索システム

2005MT036 五十住 淳　　2005MT125 八木 利夫
指導教員　河野 浩之

1　はじめに

現在，Web アーカイブ分野ではアーカイブをどのようにして保存していくかに重点が置かれ，様々な研究がなされている．しかし，保存したアーカイブから有益な情報を効率良く取得する試みについてはあまりなされていない．一方，サーチエンジン分野では有益な情報を効率良く取得するための試みとして，クラスタリングについての様々な研究がなされている．そこで，我々は，このような問題点を解決するためにアーカイブシステムにクラスタリング技術を実装することで，よりユーザーにとって効率のよい情報の取得につながると考えた．

2　Web アーカイブとセマンティッククラスタリングに関する先行研究

本章では，世界各国で行われている Web アーカイブに関する取り組みとセマンティッククラスタリングに関する先行研究について述べる．

2.1　Web アーカイブ

最近では，ますます Web 上のデータは膨大となり，その膨大なデータを後世に残していくための保存技術がより重要となってきた．このように膨大な量の Web 上のデータを保存し，次世代に残していくために Web アーカイブという技術が開発されている．現在，世界中で Web アーカイブが行われている．例えばアメリカにおいて，1996 年から Internet Archive 社が行っている Web アーカイビングプロジェクトでは，収集規模が 2.5PB となっており世界最大規模となっている．日本では 2002 年から WARP というプロジェクトが行われている．このように世界中で膨大なデータ集められているが，収集されたアーカイブから特定のアーカイブを絞って抽出して参照するとき，望んでいる情報と合致したものがうまく抽出されないという問題がある．

2.2　セマンティッククラスタリングに関する先行研究

クラスタリングとは，似ているデータ同士をひとつのグループにまとめ，ひとつの大きなデータの集合を小さなデータの集合に分類することである．効果的なクラスタリングをするには，キーワードとの意味的な関係を持つ必要がある．そこで，セマンティッククラスタリングという技術がなされている．通常のクラスタリングにセマンティックの概念を加え，意味的に重みづけをすることによって，より分類精度が向上する．このセマンティッククラスタリングを用いた研究が様々な観点でなされている．その先行研究を比較したものを表 1 に示す．

文献 [4] は，F.Jing らによる研究である．この論文では，画像検索に関してセマンティッククラスタリングを用いている．文献 [1] は H.Luo らによる研究である．この論文では，報道番組の動画に関してセマンティック分類の概念を用いている．文献 [2] は L.AlSumit らによる研究である．この論文では，テキストドキュメントに関してセマンティック分類の概念を使用している．

3　LSI+k-means クラスタリングを用いたアーカイブ検索システムの提案

本研究で提案するアーカイブ検索システムについて述べる．

3.1　文書クラスタリング

我々はテキストを対象にクラスタリングを行うことにする．文書クラスタリングの研究課題に，計算量の問題がある．計算量の問題を解決する有力な方法の 1 つに，ベクトル及びクラスタベクトルの次元を減らす (すなわち語を減らす) ことがある．このように，語を減らすことによってそれだけ共有する文書の組数が少なくなるので，計算量が改善され，類似度の計算が早くなる．この次元を圧縮する方法には以下のようなものがある．

1．何らかの語の重みに従って，語を選択する方法

2．Latent Semantic Indexing (LSI) を応用する方法

本研究では，クラスタリングを行う際に，計算量の多さを考慮して，LSI を応用する方法を選択し，応用する際のアルゴリズムとして k-means 法を採用することにする．

3.2　Latent Semantic Indexing (LSI)

LSI とは，S.Deerwester らによって提案された情報検索モデルの手法の 1 つである．LSI は情報検索の向上のために，文書と用語からなる行列に対して特異値分解 (Singular Value Decomposition:SVD) を行うことで，文書空間の次元を圧縮する手法である．次元を圧縮することで，類似度計算にかかる計算時間を圧縮することができる．

3.3　k-means 法

k-means 法は，クラスターの個数をあらかじめ指定し，個体を k 個のクラスターに分割し，そのクラスター内部で中心をとり再度クラスター分割しなおすということを繰り返す手法である．分割の基準として，クラスターの中心と各個体との間のユークリッド距離の 2 乗を用いる．

3.4　アーカイブ検索システムの概要

現在，アーカイブから有益な情報を効率良く取得する試みがあまりなされていない．2 章でも紹介した Inter-

表 1　セマンティッククラスタリングに関する先行研究

論文	観点	長所	短所
文献 [4]	画像	・画像どうしの関連性があることによりユーザーの望む画像の発見が容易	・クラスタリングが 1 度しか行われないため関連画像の階層構造が未提供
文献 [1]	動画 (報道番組)	・ユーザーにとって興味がある重要な報道ビデオを発見可能 ・全体の出来事を把握するのに効果的	・ある報道の詳細を知るには不向き
文献 [2]	テキスト ドキュメント	・セマンティック性を加えているため，既存のクラスタリング手法より精度が向上	・分類精度についてさらなる実験が必要

net Archive では，Web コンテンツおいて過去のページを閲覧できるという利点があるが，キーワード検索機能がなく，意図した情報を取得することが困難である．また，Web コンテンツ以外のデータにおいては，検索が可能であるが，日本語のコンテンツが利用できない．

これらの問題を解決するため，本研究では，アーカイブシステムに日本語にも対応した Namazu を用いて，クラスタリング技術を導入し，効率の良い情報の取得を目指す．なお，クラスタリングには，Latent Semantic Indexing (LSI) と k-means 法を組み合わせた手法を用いる．図 1 は，本研究で提案するアーカイブ検索システムの概要である．

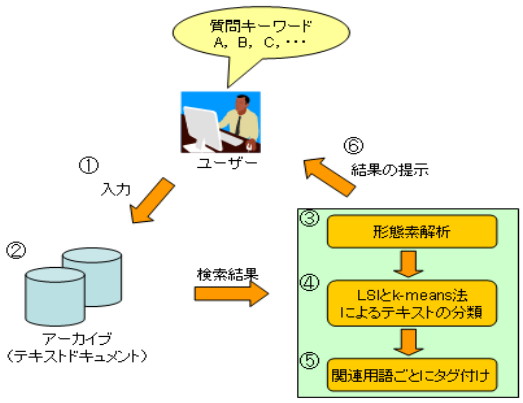

図 1　アーカイブ検索システムの概要

図 1 の説明を以下に示す．

1. ユーザーが質問キーワードを入力し，検索を行う
2. 入力されたキーワードを基にアーカイブからテキストドキュメントを検索する
3. 検索されたテキストドキュメントを形態素解析し，用語ごとに分割する．なお，形態素解析には，茶筌を用いる
4. 形態素解析された用語を基にドキュメント内の用語の類似度を Latent Semantic Indexing (LSI) を用いて求める．LSI によって重みづけされた類似性を踏まえ，テキストドキュメントを k-means 法により，クラスタリングを行う
5. 質問キーワードにヒットした文章中に頻出する用語をタグ付けする．タグ付けをすることにより，関連用語でも検索が可能になる
6. 検索結果をユーザーに提示する

4　Namazu の構築

本章では，アーカイブ検索システムとして選択した Namazu の構築とその関連ツールについて述べる．

4.1　Namazu の動作環境

Namazu は手軽に使えることを第 1 に目指した日本語全文検索システムである [3]．Namazu は CGI として動作させることにより，小中規模の WWW 全文検索システムを構築することができるほか，コマンドラインや Emacs 上で利用するといった個人用途にも使用することができるフリーフェアである．Namazu はインデックスという索引ファイルを用いているため高速な検索が可能となっている．本研究で用いた Namazu の動作環境は，表 2 に記述した通りである．

4.2　PHP・Perl からの Namazu の利用

Namazu は，CGI 上で動かすことができるが，CGI 版ではカスタマイズを行うにあたって限界がある．そこで，Namazu をカスタマイズする幅を広げるために，PHP から Namazu を動かす方法と Perl から Namazu を動かす Search::Namazu を扱う方法を本研究では試みた．ここでは，Perl から Namazu を動かす方法について述べる．

Search::Namazu は，Perl モジュールの 1 つであり，Perl のスクリプトから Namazu による検索が行うことができる．Search::Namazu を動作させるには，Namazu がインストールされている環境が必要であり，表 2 の実行環境を基に行った．Search::Namazu では，基本的なインタフェースとして，Search::Namazu::Search という関数が用意されている．以下の図 2 のように，関数を用意し，キーワードとインデックスの場所を宣言することによって検索が実行でき，検索結果として Search::Namazu::Result オブジェクトの配列を返す．Search::Namazu::Result オブジェクトに返される情報としては，タイトル，著者，日付，要約，URI，スコア，

表 2　Namazu の動作環境

Perl	File::MMagic	GNU gettext	nkf	KAKASI	Text-KAKASI	ChaSen	Text-ChaSen	Apache
5.8.8	1.23	0.16	2.0.5	2.3.4	2.0.4	2.3.3	1.0.4	2.2.3

ランク，ファイルサイズの 8 つである．

```
$query="keyword";
@result=Search::Namazu::Search(
  index = ['/var/www/index/site1'],
  lang => 'ja',
  query=> 'query',
);
```

図 2　Perl から Namazu 関数を呼び出す
サンプルプログラムの一部

4.3　形態素解析について

Namazu の動作環境の中に含まれていた KAKASI と茶筌 (ChaSen) は，形態素解析を行うツールである．KAKASI と茶筌はそれぞれの特徴があり，一長一短である．ここではそれぞれの特徴について述べ，比較していく．

まず KAKASI について述べる．KAKASI は，日本語の漢字仮名交じり文をひらがな文やローマ字文に変換するプログラムと辞書の総称である．単語ごとにわかち書きができるため，形態素解析エンジンとして Namazu などの全文検索システムと組み合わせて使われることが多い．ただし，KAKASI は，わかち書きソフトとは呼べるが，品詞情報の抽出を行うことができないため，完璧な形態素解析ソフトとは呼べないところがある．

次に，茶筌について述べる．茶筌は形態素解析用のプログラム (ChaSen) と辞書 (ipadic) の 1 組で茶筌という形態素解析ツールの機能が実現されている．茶筌においては，品詞情報の抽出が行えるため，きめ細かい日本語処理が期待できる．ただし，英語のみで構成された文章を形態素解析する際には，英単語が 1 文字単位に解析され，全て「記号-アルファベット」として認識されてしまうという弱点がある．

一般的に，形態素解析にかかる時間は KAKASI の方が若干早い．また，KAKASI はプログラムと辞書が内在しているといった拡張性の高さや，英文の処理を行う点において茶筌よりも優れている．一方で，本格的な日本語の処理を行うには茶筌の方が優れている．よって扱うアーカイブの内容によって使いわけていくことが必要である．

5　LSI + k-means クラスタリングを用いたアーカイブ検索システムの実装

ここでは，我々が行った Namazu への実装について，詳しく述べる．

5.1　アーカイブの構築

アーカイブを構築するために Web ページからデータを収集する．収集物は文書ファイル，とりわけ PDF に限定し，250 ファイルほど収集する．収集する分野は，「河野研究室分野」の (「GIS」,「P2P」) に限定することで，検索結果の関連性がよりユーザーが望むものになっているかの判断がしやすいようにした．収集したデータをアーカイブとして構築するには，インデックスを作成するディレクトリを作成し，Namazu のインデックス作成コマンド mknmz を実行する．mknmz を実行すると，指定したディレクトリ以下に NMZ.*というファイルが作成され，検索が行えるようになる．また，一つのディレクトリに複数のインデックスを作成することもできる．

5.2　Namazu を利用したアーカイブ検索システム

本研究では，LSI + k-means プログラムを Java 言語で構築した．そこで，作成した LSI + k-means プログラムを Namazu に組み込むにあたって，Perl 言語で書かれたプログラムには Java 言語を扱う方法がいくつかあるため，本研究では，Perl から利用する Namazu(Search::Namazu) を選んだ．

5.2.1　Perl 言語への Java 言語の組み込み

Search::Namazu は，Perl 言語で動作する．よって，Perl 言語で書かれたプログラムに Java 言語を扱うには，モジュールをダウンロードし，インストールする必要がある．モジュールには，代表的なものに Inline::Java というモジュールがある．Inline::Java は，Java で提供されているパッケージを利用でき，Perl 言語で書かれたプログラムの中に直接 Java プログラムを埋め込むことができるという利点がある．本研究では，この Inline::Java を用いて，作成した LSI+k-means のプログラムを Search::Namazu に組み込んだ．

5.2.2　茶筌の設定ファイルの変更

茶筌で英単語を形態素解析した場合，1 文字ずつに分解されてしまい，英単語の解析には意味をなさなくなってしまう．また，「P2P」などのアルファベットと数字が混ざり合ってる単語も同じく 1 文字ずつに分解されてしまう．そこで，この問題を解決するために本研究では，ChaSen の設定ファイル (chasenrc) に図 3 の 2 行を追加した．

```
(連結品詞 ((記号 アルファベット)))
(COMPOSIT/_POS ((名詞 一般) (名詞 数)
               (記号 アルファベット)))
```

図 3　chasenrc に追加した行

このように設定することで，英単語が 1 つの単語として解析されるようになる．ただし，英単語においては，1 つの英単語に様々な品詞があり，文章の解析において英単語の品詞判定が難しいため，本研究では，「名詞-一般」として解析するものとした．

5.2.3 LSI + k-means クラスタリングを用いたアーカイブ検索システム

ここでは，我々が作成したプログラムの概要を述べる．本研究では，search.html と search-namazu.cgi の 2 つのプログラムを作成した．

1．search.html

検索したいキーワードを入力し，その情報を検索結果ページへ送るプログラムである．図 4 は，search.html で作成した検索を行うトップ画面である．

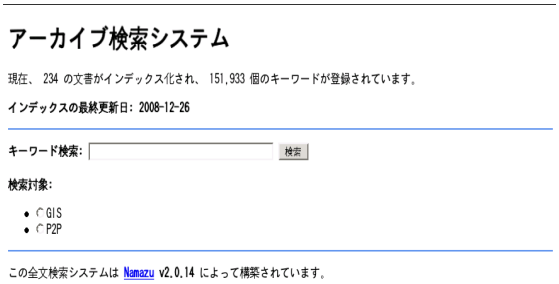

図 4　search.html で作成した検索トップ画面

まず，入力したいキーワードをテキストボックスに入力する．そして，「GIS」か「P2P」のどちらかの検索対象を選び，検索を行う．キーワードに関しては，複数キーワードを入力するとはるかに処理が遅くなる可能性があるため，1 つのキーワードのみを入力して検索を行うことに限定した．

2．search-namazu.cgi

search.html から送られてきたキーワードを基に，Namazu による検索を行い，検索結果を表示するプログラムである．図 5 は search.html から送られてきた情報を基に検索を行った検索結果画面である．

ここでの例では，「システム」というキーワードを基に検索を行っている．また，検索結果の表示以外に，文章を形態素解析した際の単語のうち，文章中に頻出する単語をキーワードに関連する単語として再度検索できるようにタグづけを行った．なお，図 5 のように，関連用語に英語と日本語のキーワードが混在する場合もあるため，英語だけでなく日本語も同時に解析し，名詞のみの抽出を行うため，品詞情報まで解析できる茶筌を用いて解析することにした．ただし，形態素解析された語句の中に関連用語として意味をなさない語句や人名などの単語も含まれることがあるため，意味をなさない語句を含めないようストップワードとして省く処理を行った．また，5.2.2 節のように設定を行うことで，関連用語において英単語が 1 つの単語として解析されていることが分かる．なお，クラスタリングを行う際のクラスタ数は，

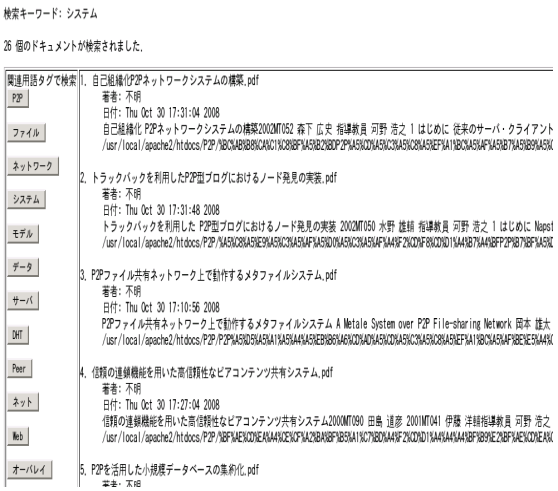

図 5　search-namazu.cgi で作成した検索結果画面

250 ファイルのドキュメント数を考慮し，手動で分類精度の実験を繰り返し試行錯誤した結果，分類精度の最も良かった 5 を選択した．

6　まとめ

本研究では，アーカイブシステムに対し，LSI + k-means クラスタリングを導入した．アーカイブシステムに Namazu を用いることにより，キーワード検索が可能になり，英語のみならず，日本語のコンテンツにも対応することができた．

今後の課題としては，扱うアーカイブの言語，内容に対応した辞書を作成して追加することや他のアーカイブ検索システムの性能評価を行い，システムの実用性を検証していくことが挙げられる．また，ドキュメント数が変化する度に最適なクラスタ数を設定し直す必要がある．

参考文献

[1] H.Luo, J.Fan, J.Yang, W.Ribarsky, S.Satoh: “Analyzing Large-Scale News Video Database to Support Knowledge Visualization and Intuitive Retrieval,” Proceedings of the IEEE Symposium on Visual Analytics Science and Technology 2007, pp.107-114, 2007.

[2] L.AlSumit，C.Domeniconi: “Local Semantic Kernels for Text Document Clustering,” SIAM International Conference on Data Mining, 2007.

[3] Namazu Project:　全文検索システム Namazu, http://www.namazu.org/(accessed 2008.12.29).

[4] S.Wang, F.Jing, J.He, Q.Du, L.Zhang: “IGroup: presenting web image search results in semantic clusters,” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp.587-596, 2007.