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

2007－DBS－143（54）
2007／7／3

歴史関連ドキュメントを対象とした多元的情報可視化システムの実現

井上 勝哉† 佐々木史織†† 清木 康†††

† 慶応義塾大学総合政策学部 〒252-0011 神奈川県藤沢市遠藤 5322
†† 慶應義塾大学大学院政策・メディア研究科 〒252-0011 神奈川県藤沢市遠藤 5322
††† 慶応義塾大学環境情報学部 〒252-0011 神奈川県藤沢市遠藤 5322
E-mail: †inokatsu@mdbl.sfc.keio.ac.jp, ††sashiori@mdbl.sfc.keio.ac.jp, †††kiyoki@mdbl.sfc.keio.ac.jp

あらまし 本稿では、メタデータが付与されていない歴史に関する文書データ (歴史関連ドキュメント) 群を対象に、時空間情報による文書分類と、地理情報データへの変換及び可視化を行い、一元的な文字集合のテキスト媒体から多元的な情報表現を持つ媒体へと変換する歴史情報源表現システムの実現方式を示す。本方式により実現されたシステムを用いることにより、システム利用者は、文書に直接明示されていない時代分類情報と、時代ごとの歴史的事項と地理的な場所との対応を把握することが可能となる。本稿では、実際の高校生向け日本史用語集を用いて時代分類機能のためのメタデータベースを作成し、処理対象の歴史文書として web 上のコミュニティベース百科事典の日本史の項目を用いた実験を行い、本方式の実現可能性を示す。
キーワード 歴史文書, データ表現, 文書分類, 地理情報, 可視化

A Visualization System for Realizing Multiple Views for Historical Documents

Katsuya INOUE†, Shiori SASAKI††, and Yasushi KIYOKI†††

† Faculty of Policy Management, Keio University
†† Graduate School of Media and Governance, Keio University
††† Faculty of Environmental Information, Keio University
E-mail: †inokatsu@mdbl.sfc.keio.ac.jp, ††sashiori@mdbl.sfc.keio.ac.jp, †††kiyoki@mdbl.sfc.keio.ac.jp

Abstract In this paper, we present an implementation method of a visualization system for realizing multiple views for historical documents. This system enables document classification by periodic information and visualize geographic information from simple plain texts about history. By using the implemented system by our method, users can grasp visually where historical events occurred or can speculate an era the processed document describe. To examine the availability of the system, we performed several experiments using documents on the free encyclopedia on the web which describe Japanese historical terms and events.
Key words historical documents, data representation, document classification, geographical information, visualization

1. はじめに

近年、世に出る文書の殆どは計算機で何らかの処理が可能な形式になっており、文献 [1] によれば、WWW 上の文書データ量は 200TB を超えていると言われている。歴史に関する文書も WWW 上に散見される。文書の検索を目的とした計算機システムも一般に普及しており、歴史用語を検索呼出し語として WWW 検索システムに入力すれば、検索呼出し語にパターンマッチした webpage のリストが何らかの基準でインデクシングされて出力される。

しかし、歴史に関する文書データ (歴史関連ドキュメント) の読み手にとっては時代分類と地理情報が重要であるにも関わらず、既存の文書検索システムでは、複数の時代・場所にまたがって関連する文書などの時代情報の適切な分類がなされず、分類された結果と地理情報を直観的に利用者に提示する可視化システムも実現されていない。

そこで本稿では、歴史関連ドキュメント群を対象に、各文書データに含まれる時代別の特徴語 (歴史上の事象・人物名・地名) を基に、ドキュメントの時代分類と二次元地図上への空間配置を自動化し、歴史文書データの読み手に対して、より付加

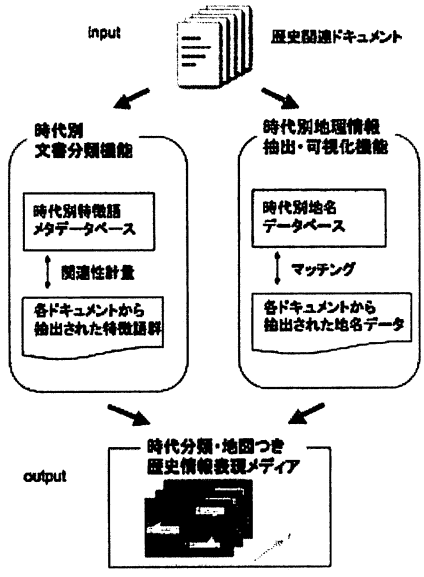


図 1　システム全体の構成図
Fig. 1　An Overall System Architecture

価値の高い情報表現を生成する新しい歴史情報源表現システムの実現方式を示す。

第一に、歴史に関する学術文献を元データとして、各時代を特徴づける歴史上の事象・人物・地理情報に関するメタデータベースを作成する。第二に、各文書データに含まれる歴史上の事象・人物名などの情報を各時代を表す特徴語として抽出し、時代による文書分類の判定材料とする。

第三に、各文書データに含まれる地理情報を二次元地図上へ配置するための処理機構を作成する。

文書に直接明示されていない時代分類情報と地理情報を得ることで、システム利用者は歴史に関する文書からより多元的な情報を得ることが可能である。

2. システム概要

本システムの全体の構成図は図 1 の様に示される。

本システムの主要部分は、時代による文書分類 (document classification) を行う時代別文書分類機能と、地理情報の可視化を行う地理情報抽出可視化機能の 2 つから構成される。

まず、歴史に関する学術文献等から時代分類器に於いて使用するメタデータを作成する。

次に、時代別文書分類機能に於けるメタデータ中の用語の中で地名、地理情報であると判定されうる用語群に対して、緯度経度情報を付加したデータ対応表を作成しておく。

処理対象歴史ドキュメントに対して、時代別文書分類機能で使用するメタデータ中で、特定の時代に割り振られた用語がどれくらい出現するかを計算し、処理対象歴史ドキュメントの時代分類情報が決定される。

地理情報抽出可視化機能に於いても、処理対象歴史ドキュメントに対して、作成済みの緯度経度情報と地名データの対応表中に記された地名データがあれば、その組み合わせを基に外部 Geocoding Engine に対して問い合わせを行い、処理対象歴史ドキュメントに記された地名情報の地図を生成する。

3. 時代分類機能の実現方式

歴史ドキュメントの時代分類機能を以下の手順で構築する。

- 時代分類歴史用語メタデータベースの準備
- 関連性計量機構の構築

3.1 時代分類歴史用語メタデータの準備

時代分類を行うにあたり、日本史の学術用語集である文献 [3] を用い、時代分類歴史用語メタデータを作成する。時代分類歴史用語メタデータは、以下のように 2 つの層で構成される。

（1） A 層:direct definition layer

A 層では、個々の時代ごとの特徴を直接表す歴史的事象や人物の名前を直接的に示す用語の集合で、ある 1 つの時代分類を特徴付ける。ある時代 X を特徴付ける用語が n 種類あるとすれば、以下のように時代 X の特徴が表される。

$$\text{時代 } X := (\text{用語 } 1, \text{用語 } 2, \text{用語 } 3, \cdots, \text{用語 } n) \quad (1)$$

A 層に於いて、時代を直接的に特徴付けている用語を、第 1 次特徴語と定義する。

（2） B 層:indirect definition layer

B 層では、第 1 次特徴語を、文献 [3] 内にて解説している文中に出現する周辺的な歴史用語や、一般名詞を用いて第 1 次特徴語を特徴付ける。第 1 次特徴語 Y を特徴付ける用語が m 種類あるとすれば、以下のように第 1 次特徴語 Y の特徴が表される。

$$\text{第 1 次特徴語 } Y := (\text{用語}\alpha_1, \text{用語}\alpha_2, \text{用語}\alpha_3, \cdots, \text{用語}\alpha_m) \quad (2)$$

B 層に於いて、第 1 次特徴語を特徴付けている用語を、第 2 次特徴語と定義する。第 2 次特徴語の集合には、歴史用語だけでなく一般の用語も含まれており、時代をより直接的に特徴付ける歴史用語が含まれていない文書に対しても、時代分類の計量の対象とすることができる。

3.2 関連性計量機構の構築

上述した A 層と B 層の二層構造を取る時代分類別歴史用語メタデータを用いて、処理対象歴史ドキュメントの、時代分類に対する関連性計量を行う。

3.2.1 第 1 次特徴語による関連性計量

（1） メタデータ A 層の第 1 次特徴語の集合を作成

用意したメタデータの A 層に含まれている第 1 次特徴語を全て列挙し、重複した要素は除去して冗長性を排除し、単一の集合の要素とする。

（2） 包括第 1 次特徴語 vector 空間を作成

重複を除去した集合の要素を vector の要素とする包括第 1 次特徴語 vector 空間を作成する。メタデータ中で定義する時代が時代 α , 時代 β, 時代 γ の三種類に限定されているものと

仮定する場合、A 層のメタデータは以下のように記述できる。

$$\text{時代}\alpha := (\text{特徴語 } A_1, \text{特徴語 } A_2, \text{特徴語 } A_3, \cdots, \text{特徴語 } A_n) \quad (3)$$

$$\text{時代}\beta := (\text{特徴語 } B_1, \text{特徴語 } B_2, \text{特徴語 } B_3, \cdots, \text{特徴語 } B_m) \quad (4)$$

$$\text{時代}\gamma := (\text{特徴語 } C_1, \text{特徴語 } C_2, \text{特徴語 } C_3, \cdots, \text{特徴語 } C_l) \quad (5)$$

上記のように A 層が構成されている場合の特徴語の重複が o 件であると仮定すると、包括第 1 次特徴語 vector 空間は (n + m + l − o) 次元となる。

$$\text{包括第 1 次特徴語 } vector \text{ 空間} = (\chi_1, \chi_2, \chi_3, \cdots, \chi_{n+m+l-o}) \quad (6)$$

（3） 時代別に第 1 次特徴語 vector データを作成

包括第 1 次特徴語 vector 空間と同じ次元数を持ち、各成分が時代別の第 1 次特徴語に対応した時代別の第 1 次特徴語 vector データを作成する。計量機能の時代関連性計算と、その計算結果を正規化することを考慮し、ある時代の包括第 1 次特徴語 vector 空間のある特徴が含まれていれば、第 1 次特徴語 vector の成分の値を “1” とし、含まれていなければ “0” とする。

	特徴語	特徴語	特徴語	···	特徴語
時代 α	0	0	1		0
時代 β	0	1	1		0
時代 γ	1	0	0		1

（4） 処理対象歴史文書の第 1 次特徴語 vector を作成

処理対象となっている歴史文書を形態素解析し、単語以外の記号文字や、html 要素の開始・終了 tag などの不要データを除去した上で、分かち書き済みの文書データに加工する。

加工した文書データに包括第 1 次特徴語 vector 空間の各成分が示す歴史用語が含まれているかどうかを検知した上で、包括第 1 次特徴語 vector 空間と同じ次元数を持ち、各成分が第 1 次特徴語に対応した処理対象歴史文書の第 1 次特徴語 vector データを作成する。

対応する包括第 1 次特徴語 vector 空間の特徴が含まれていれば、処理対象歴史文書の vector の成分の値を “1” とし、含まれていなければ “0” とする。

（5） 二種類の vector 同士の内積計算

ある歴史文書と時代 α の関連性を、時代を定義している第 1 次特徴語によって計算する場合、以下のような内積計算を行う。ここで、

時代 α の第 1 次特徴語 vector データを $\alpha_n = (A_{x1}, A_{x2}, A_{x3}, \cdots, A_{xn})$, ある歴史文書 nu の第 1 次特徴語 vector データを $\nu_n = (A_{y1}, A_{y2}, A_{y3}, \cdots, A_{yn})$ とすると、その内積 $\delta(\alpha_x, \nu_y)$ は

$$\delta(\alpha_x, \nu_y) = \sum_{i=1}^{n} A_{xi} \cdot A_{yi} \quad (7)$$

となる。前述したように、時代別に定義した第 1 次特徴語 vector データも、処理対象歴史文書の第 1 次特徴語 vector データも、各成分の値は 0 か 1 のどちらかである。したがって、内積の値は、処理対象歴史文書の中に、関連性を計量したい時代を特徴付ける歴史用語が重複をカウントせず幾つ存在するかを計算することとほぼ同義である。

（6） 計量値の正規化

A 層のメタデータの中でも、第 1 次特徴語の登録してある数は、時代分類毎に異なっているため、第 1 次特徴語を多く登録してある時代の関連性計量値は大きくなりやすい。よって、上記のように内積を計算したのみで文書を分類する場合、時代分類ごとの第 1 次特徴語の登録数で、内積計算の結果を割った値で判定する。冗長性を排除した時代 α の第 1 次特徴語の登録数を m とすると、ある時代文書 ν と時代 α の関連性計量値は

$$(\sum_{i=1}^{n} A_{xi} \cdot A_{yi}) \div m \quad (8)$$

となる。

A 層のメタデータが ψ 種類の時代分類を示すものから構成されている場合、処理対象文書と時代との関連性計量を ψ 回行い、最も関連性計量値が高かった時代が、第 1 次特徴語による時代分類の結果となる。

処理対象ドキュメント中に、歴史用語が明示的に記されていることが明白な場合、第 1 次特徴語による関連性計量が有効であると予想される。

3.2.2 第 2 次特徴語による関連性計量

第 1 次特徴語 Y を特徴付ける第 2 次特徴語が m 種類あるとする場合、以下のように第 1 次特徴語 Y の特徴が表される。(第 2 次特徴語を “2F” と表す。)

$$\text{第 1 次特徴語 } Y := (2F\alpha_1, 2F\alpha_2, 2F\alpha_3, \cdots, 2F\alpha_m) \quad (9)$$

（1） メタデータ B 層の第 2 次特徴語の集合を作成

用意したメタデータの B 層に含まれている第 2 次特徴語を全て列挙し、重複した要素は除去して冗長性を排除し、単一の集合の要素とする。

（2） 包括第 2 次特徴語 vector 空間を作成

集合の要素を vector の要素とする包括第 2 次特徴語 vector 空間を作成する。A 層メタデータ中で定義した時代 α の第 1 次特徴語 vector が、時代α := (特徴語 A_1, 特徴語 A_2, 特徴語 A_3, ・・・, 特徴語 A_n), だと仮定し、第 1 次特徴語を “1F” 第 2 次特徴語を “2F” と表すとき、時代 α を示す B 層のメタデータは以下のように示される。

$1FA_1$	$2FA_{1-1}$	$2FA_{1-2}$	$2FA_{1-3}$	···	$2FA_{1-\epsilon}$
$1FA_2$	$2FA_{2-1}$	$2FA_{2-2}$	$2FA_{2-3}$	···	$2FA_{2-\rho}$
$1FA_3$	$2FA_{3-1}$	$2FA_{3-2}$	$2FA_{3-3}$	···	$2FA_{3-\mu}$
中略	···	···	···	···	···
$1FA_n$	$2FA_{n-1}$	$2FA_{n-2}$	$2FA_{n-3}$	···	$2FA_{n-\kappa}$

また、A 層で包括第一次特徴語 vector 空間が特徴付けた時代が ϑ 個あり、B 層中で、1 つの第 2 次特徴語に付き、特徴付けしている第 2 次特徴語の重複が $\zeta_{1\vartheta}$ 件であると仮定すると、

包括第 2 次特徴語 vector 空間は $(\epsilon_1 + \rho_1 + \mu_1 \cdots + \kappa_1 - \zeta_1) +$

$(\epsilon_2+\rho_2+\mu_2\cdots+\kappa_2-\zeta_2)+(\epsilon_3+\rho_3+\mu_3\cdots+\kappa_3-\zeta_3)+\cdots$ 中略 $\cdots+(\epsilon_\vartheta+\rho_\vartheta+\mu_\vartheta\cdots+\kappa_\vartheta-\zeta_\vartheta)$ 次元となる。

（3） 時代別に第 2 次特徴語 vector データを作成

包括第 2 次特徴語 vector 空間と同じ次元数を持ち、各成分が第 2 次特徴語に対応した時代別の第 2 次特徴語 vector データを作成する。計量機能の時代関連性計算と、その計算結果を正規化することを考慮し、ある時代の包括第 2 次特徴語 vector 空間のある特徴が含まれていれば、時代別の第 2 次特徴語 vector の成分の値を "1" とし、含まれていなければ "0" とする。包括第 2 次特徴語 vector 空間の次元数を Z とし、第 2 次特徴語を "2F" と表すと、

	$2F_1$	$2F_2$	$2F_3$	$\cdots$	$2F_z$
時代 α	0	0	1		0
時代 β	0	1	1		0
時代 γ	1	0	0		1

のように第 2 次特徴語 vector データが定義される。

（4） 処理対象歴史文書の第 2 次特徴語 vector を作成

処理対象となっている歴史文書を形態素解析し、単語以外の記号文字や、html 要素の開始・終了 tag などの不要データを除去した上で、分かち書き済みの文書データに加工する。

加工した文書データに包括第 2 次特徴語 vector 空間の各成分が示す歴史用語が含まれているかどうか検索し、そして包括第 2 次特徴語 vector 空間と同じ次元数を持ち、各成分が第 2 次特徴語に対応した処理対象歴史文書の第 2 次特徴語 vector データを作成する。

対応する包括第 2 次特徴語 vector 空間の特徴が含まれていれば、処理対象歴史文書の vector の成分の値を "1" とし、含まれていなければ "0" とする。

（5） 二種類の vector 同士の内積を計算

ある歴史文書と時代 β の関連性を、時代を定義している第 2 次特徴語によって計算する場合、第 1 次特徴語による計算と同様に、まずは以下のような内積計算を行う。包括第 2 次特徴語 vector 空間の次元数を ζ と仮定すると、

時代 β の第 2 次特徴語 vector データを $\beta_n = (A_{x1}, A_{x2}, A_{x3}, \cdots, A_{x\zeta})$, ある歴史文書 nu の第 1 次特徴語 vector データを $\nu_n = (A_{y1}, A_{y2}, A_{y3}, \cdots, A_{y\zeta})$ と表すと、その内積 $\delta(\beta_x, \nu_y)$ は

$$\delta(\beta_x, \nu_y) = \sum_{i=1}^{\zeta} A_{xi} \cdot A_{yi} \tag{10}$$

となる。3.2.1 の第 1 次特徴語 vector と同様に、時代別に定義した第 2 次特徴語 vector データも、処理対象歴史文書の第 2 次特徴語 vector データも、各成分の値は 0 か 1 のどちらかである。よって、内積の値は、処理対象歴史文書の中に、関連性を計量したい時代を特徴付ける歴史用語が重複をカウントせず幾つ存在するかということを計算するのと同じである。

（6） 計量値の正規化

A 層と同じく B 層のメタデータの中でも、第 2 次特徴語の登録してある数は、時代分類毎に異なっているので、第 2 次特徴語を多く登録してある時代の関連性計量値は大きくなりやすい。よって、上記のように内積を計算しただけで文書を分類すると、問題が生ずるので、時代分類ごとの第 2 次特徴語の登録数で、内積計算の結果を割った値で判定する。冗長性を排除した A 層で時代 β を特徴付けている第 1 次特徴語の集合を取り、その要素数が m だとすると、B 層でその集合の要素の一つ一つである用語を特徴付ける第 2 次特徴語を m 個の第 1 次特徴語ごとに数え上げていき、重複を排除した数が n だとする。このとき、ある時代文書 ν と時代 β の関連性計量値は

$$(\sum_{i=1}^{\zeta} A_{xi} \cdot A_{yi}) \div n \tag{11}$$

となる。

A 層のメタデータが ζ 種類の時代分類を示すものから構成されている場合、処理対象文書と時代との関連性計量を ζ 回行い、最も関連性計量値が高かった時代が、第 2 次特徴語による時代分類の結果となる。

B 層で用いられる第 2 次特徴語は一般的な言葉や、広範な時代にまたがった語、より専門的な歴史用語である為、処理対象ドキュメント中に、ある時代を示唆する歴史用語が明示的に記されていない場合、第 2 次特徴語による関連性計量が有効であると予測される。

4. 地理情報抽出・可視化機能の実現方式

本システムに於ける地理情報抽出・可視化機能は、時代別地理情報データベースと、緯度経度データ写像機構から構成される。時代別地理情報データベースは、処理対象歴史文書中の地名情報を抜き出し、メタデータ中の地名と緯度経度情報の対応表を参照し、地名情報と緯度経度情報の順序対 (ordered pair) が要素である集合を生成する。そして、得られた地名情報と緯度経度情報の順序対の集合に従って、緯度経度データ写像機構がデジタルマップを生成し可視化を行う。

4.1 時代別地理情報データベースの構築

4.1.1 時代別地理情報メタデータの準備

前章で定義した時代分類機能の為のメタデータである全ての第 1 次特徴語と第 2 次特徴語の中から、地名情報だと解釈されるものを列挙する。

列挙した地名情報に対する緯度経度情報を調べ上げ、地名情報と緯度経度情報の順序対 (ordered pair) を集合の要素とした地名・緯度経度情報の対応表を作成する。

$$\begin{aligned}&((\text{地名}_1, \text{緯度}_1, \text{経度}_1), (\text{地名}_2, \text{緯度}_2, \text{経度}_2),\\ &(\text{地名}_3, \text{緯度}_3, \text{経度}_3), \cdots, (\text{地名}_n, \text{緯度}_n, \text{経度}_n))\end{aligned} \tag{12}$$

4.1.2 地名情報の処理

処理対象の歴史関連文書から時代別の地理情報を生成する為に、前節 3.1 および 3.2 で説明した時代分類機能で処理した結果の時代分類情報を活用する。時代分類の結果が時代 α であったと仮定する。

文書分類機能を用いた時代分類の結果が時代 α であるから、時代 α を特徴付けている 1 次特徴語と 2 次特徴語が地名情報

としてそのまま使える場合があることを利用する。

（1） 前章で言及した時代分類の為のメタデータである A 層と B 層を参照し、時代 α を特徴付ける第 1 次特徴語と第 2 次特徴語を全て列挙する。

（2） 列挙した第 1 次特徴語と第 2 次特徴語の集合と、地名・緯度経度情報の対応表の中の地名で、共通の要素となる地名があれば、緯度経度情報も含めて列挙する。

（3） 処理対象文書中に出現する "時代 α に於ける地名と緯度経度情報の順序対" が要素の集合が完成する。

さらに、時代 α を特徴付けている地名ではあるが、処理対象歴史文書中に出現しない地名を、「文書中にはないが同じ時代の重要地」という目的で可視化を行いたいというケースの場合は、次のプロセスに従って処理を行う。

（1） 前章で言及した時代分類の為のメタデータである A 層と B 層を参照し、時代 α を特徴付ける第 1 次特徴語と第 2 次特徴語を全て列挙する。

（2） "列挙した第 1 次特徴語と第 2 次特徴語の集合" と、"地名・緯度経度情報の対応表の中の地名で、処理対象文書には出現せず、時代分類の結果の時代を特徴付けている第 1 次特徴語と第 2 次特徴語の集合の中にある地名データ" と共通の要素ではない地名があれば、緯度経度情報も含めて列挙する。

（3） 処理対象文書中に出現する "時代 α に於ける地名と緯度経度情報の順序対" が要素の集合が完成する。

最後に、時代 α に於ける "地名と緯度経度情報の順序対" が要素の集合を、時代分類情報と共に緯度経度写像機構に渡す。

4.2 緯度経度データ写像機構の構築

4.1 のプロセスを経て、地名と緯度経度の組み合わせを得た後、処理対象文書の plain text 部分を加工したものを付加して、外部の geocoding system を利用して可視化を行う。本実装では Google Maps API を使用する。

5. 時代分類機能の実験

本システムの時代分類機能の適合率 (precision) を計測する為に、以下の実験を行った。

5.1 実験環境

例として、コミュニティベースの百科事典 [2] で "旧石器時代" の項目にある「土器」を時代分類機能で処理すると、次のような情報が出力される。

計量手法	分類結果	内積	時代の特徴語数
第 1 次特徴語	旧石器 (無効)	0	10
第 2 次特徴語	縄文時代	3	106

文献 [2] では「土器」の場合、"旧石器時代" という時代分類なので、下の出力の分類結果では第 1 特徴語による分類も、第 2 特徴語による分類も適合しない結果となる。

ここでの "分類結果" は、それぞれの計量手法で内積 (式 7,10) を計量する時代の特徴語数の数で割ったもの (式 8,11) が最高値であった時代を列挙している。ただし、メタデータが用意されている時代の内積が 0 であった場合、無効な分類となる。

また、本稿の実験では、メタデータの準備の関係で、"旧石器時代" から "奈良時代" までの 6 つの時代分類にのみ対応している。

（1） web 上のコミュニティベース百科事典 [2] の日本史の分類項目にある見出し語の page を、処理対象歴史ドキュメント群とする。

（2） 1 つの時代につき 10 つの用語、6 つの時代で 60 の用語に関して時代分類機能にかけ、時代分類の適合率を調べる。

5.2 時代分類機能の実験結果および考察

第 1 次特徴語を "1F" 第 2 次特徴語を "2F" と表し、文献 [2] に於ける 6 つの時代項目の中からそれぞれ時代ごとに 10 種類の見出し語を選び、合計 60 の用語に関して時代分類機能を試した結果を以下の表にまとめた。(省略表記「適」… 正解と分類結果が適合、「不」… 正解と分類結果が不適合、「失」… 分類失敗 (分類不能、内積の値が 0))

	1F 適	1F 不	1F 失	2F 適	2F 不	2F 失
旧石器	6	1	3	4	6	0
縄文	6	4	0	9	1	0
弥生	6	4	0	7	3	0
古墳	9	0	1	10	0	0
飛鳥	6	4	0	2	8	0
奈良	6	4	0	2	8	0

この実験結果から以下のことが読み取れる。

- 第 2 次特徴語分類は語数が多く、一般語も多く定義してあるので、殆ど "分類失敗 (内積値が 0)" にならない。
- 飛鳥時代を古墳時代であると間違って分類されることが多いが、これは飛鳥時代と古墳時代は時間軸情報として重複しており、定義者の主観次第でどちらにも分類可能な時期があったからである。
- 旧石器時代などの、A 層のメタデータも B 層のメタデータも乏しい時代分類の場合、第 1 次特徴語による時代分類の方が精度が高くなる

6. 地理情報の可視化システムの実装実験

6.1 実験環境

（1） 時代分類機能の第 1 次特徴語の集合と第 2 次特徴語の集合から地名・緯度経度情報の対応表は作成済みであると仮定する。この対応表を "地名・緯度経度対応表 γ" とする

（2） web 上のコミュニティベース百科事典 [2] の日本史の分類項目にある見出し語 "関東ローム層" の page を地理情報抽出・可視化機構の処理対象とする。

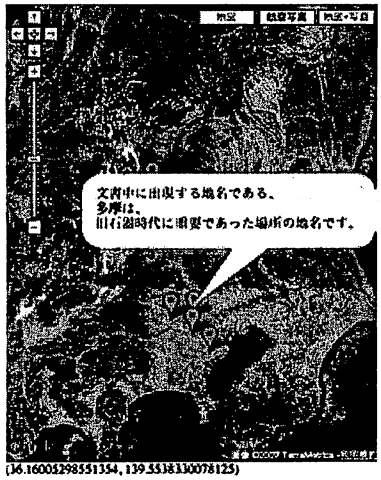
6.2 実験の過程

まず、処理対象歴史文書である「関東ローム層」の page を本システムの時代分類機能にかける。すると、以下のような分類結果になる。

計量手法	分類結果	内積	時代の特徴語数
第 1 次特徴語	旧石器 (無効)	0	10
第 2 次特徴語	旧石器	6	51

第 1 次特徴語による時代分類には失敗しているが、第 2 次特徴語による時代分類の結果から、処理対象歴史文書の示す時代は "旧石器時代" であるとする。

次に、処理対象歴史文書の時代分類が "旧石器時代" 決定し



(36.160052985551354, 139.55383300781125)

図 2　地理情報可視化システム実装実験結果 1

(36.160052985551354, 139.37805175781125)

図 3　地理情報可視化システム実装実験結果 2

たので、時代分類機能のメタデータの A 層中にて旧石器時代を特徴付ける第 1 特徴語の集合と、それらの第 1 第特徴語を B 層で特徴付けている第 2 特徴語の集合の和集合を生成する。これを集合 α とする。

$$集合\alpha = (洪積世, 石器時代, マンモス, \cdots, 群馬県) \qquad (13)$$

これを、あらかじめ作成してある、“地名・緯度経度対応表 γ” の地名の部分だけを抜き出した集合を集合 η とし、集合 η 集合 α との積集合を取ると、以下のような集合 δ が生成される。この集合 δ の要素が「処理対象歴史文書の示す時代に於ける、処理対象歴史文書にされている重要な地名データ」である。

$$集合\delta = (多摩, 下末吉, 武蔵野, 立川) \qquad (14)$$

集合 η から集合 δ に属する要素を間引いて得られる集合を、集合 β とする。この集合 β は、

$$集合\beta = (関東, 北海道, シベリア, 朝鮮半島, 角石, 笠懸, 新田, 群馬)$$

となり、この集合 β の要素が「処理対象歴史文書の示す時代に於ける、処理対象歴史文書に記述されていないが重要な地名データ」となる。これら集合 δ と集合 β を緯度経度写像機構に引き渡す。

6.3　実験の結果

以下に、集合 δ と集合 β を “地名・緯度経度対応表 γ” を参照し、地名と緯度経度の順序対を要素とする集合に変換し、緯度経度写像機構に引き渡した結果を以下に図 2 と図 3 に示した。図 2 は集合 δ の要素である “処理対象文書中に存在する地名” に焦点を合わせた可視化結果である。図 3 は集合 β の要素である “処理対象文書中に存在しない地名” に焦点を合わせた可視化結果である。この可視化結果を得ることで、システム利用者は、時代分類情報に基づいた地理情報を直感的に得ることが可能となる。

7.　結論と展望

本稿では、時空間情報による文書分類と、地理データへの変換および可視化機能を備えた歴史情報表現システムの実現方式について示した。また、web 上のコミュニティベース百科事典のデータを対象に実験を行うことで、歴史情報が含まれた文書の時代別分類と、地理情報の可視化について、実現可能性を示した。本稿に於いて示した方式では、文書分類のメタデータを如何に量的に充実させるか、メタデータを特徴付ける用語が適切かどうかが、分類と可視化の精度に強く影響してくる。今後はよりメタデータの量的充実・データ構造の最適化を進めていきたい。

8.　謝　　辞

本研究を進めるにあたって、実験システムの使用法などで多くの助言を頂いた慶應義塾大学大学院政策・メディア研究科の河本穣氏、倉林修一氏に感謝致します。

文　　献

[1] Peter Lyman and Hal R. Varian, How Much Information? 2003, http://www.sims.berkeley.edu:8000/research/projects/how-much-info-2003/ , 2003.

[2] Wikimedia Foundation, Inc.,『フリー百科事典 ウィキペディア日本語版』http://ja.wikipedia.org/wiki/ , 2007

[3] 中里 裕司, 必修 日本史 B 用語集 改訂版, ISBN-10: 4010353155 , 2002

[4] 執筆者不明, 五畿七道, http://www.geocities.jp/wuyongdeye/tables/5ki7dau.html

[5] Patric Henry Winston, Berthold Klaus Paul Horn, LISP 3rd Edition, Addison Wesley Publishing Company,Inc., ISBN 0-201-08319-1.

[6] Google, Inc., Google Maps API Documentations, http://www.google.com/apis/maps/documentation/index.htm