新聞記事コーパスにおける自然災害の特性と時間関係の抽出

吉岡 卓　　谷 聖一　　戸田 誠之助
日本大学文理学部　日本大学文理学部　日本大学文理学部

歴史学や社会学において自然災害が社会に及ぼす影響を考えることは重要な事である．近年，自然災害に関するデジタルアーカイブの作成が行われているが，これらは人手によって作成されており大変労力がかかる．そこで我々は新聞記事から自然災害アーカイブを作成するための計算機システムを提案する．いくつかのインデックスを定義することで，新聞記事内の自然災害を特徴付け，自然災害間の時間構造を示す．

Extraction of temporal relation by the creation of historical natural disaster archive

Suguru Yoshioka
College of Humanities and Sciences
Nihon University

Seiichi Tani
College of Humanities and Sciences
Nihon University

Seinosuke Toda
College of Humanities and Sciences
Nihon University

In historical science and social science, the influence of natural disaster upon society is a matter of great interest. In recent years, some archives are made through many hands for natural disasters, however it is inefficiency and waste. So, we suppose a computer system to create a historical natural disaster archive from newspaper articles. In order to do this analysis, we identify the occurrences in newspaper articles by some index entries, considering the affairs which are specific to natural disasters, and show the temporal relation between natural disasters.

1 はじめに

近年，情報科学の発展に伴い様々な研究分野のリソース (Archieve) を蓄積/情報発信し，新しい発見や問題解決を行う方法が広く研究されている [1, 3, 4, 6, 9, 11, 12]．多くの研究機関において，古文書，美術品，動画像による実験データなどがデジタル化されており，特に情報科学における研究成果を用いる事で，デジタルアーカイブの効果的な構築や利用が求められている．本学においても，多くの情報をアーカイブ化し高度利用を行うためのプロジェクト「デジタルアーカイブ・インフラストラクチャの構築と高度利用」が進行中である[1]．

本稿では特に自然災害に着目したデジタルアーカイブ作成システムの構築を提案する．我々のシステムは (i) 新聞記事から自然災害に関する記事を抽出し，テキストベースのアーカイブを作成する．また (ii) 作成したアーカイブに対して自然災害の時間的構造を自動的に特徴付ける [2] 事を目的とする．このことは，新聞記事上の事象の流れを現実における時間軸上にマッピングする事 [10] であり，我々の研究が社会学，歴史学の研究分野において役立つ事を示す．

次章では，本学のデジタルアーカイブプロジェクトと自然災害アーカイブの必要性について述べる．3 章では，自然言語処理の技術を基にした自然災害アーカイブの為の計算機システムを示す．4 章では，アーカイブにおける時間構造について説明し，いくつかの制約とルールを定義する事で災害間の時間関係を解析する方法を示す．最後に，本研究のまとめと今後の課題について述べる．

[1] 私立大学学術研究高度化推進事業 学術フロンティア推進事業 2003-2007 による私学助成を得て行われている．

2 デジタルアーカイブ

本章では，本学におけるデジタルアーカイブシステムについて述べる．特に自然災害を扱うアーカイブの特徴を示す．なお本稿では，特定の対象に応じて集められたデジタルデータをアーカイブ，そこに含まれる各要素をコンテンツと呼ぶこととする．

2.1 デジタルアーカイブプロジェクト

本学で進行中のプロジェクトでは，データの蓄積と検索を可能とする計算機システムが構築されている．様々な分野の研究者によって，古文書，デジタル地図，実験データなどの貴重資料がデジタル化され蓄積されている．これらのコンテンツはインターネットを介して多くの人々が利用可能となっている[2]．我々はデジタル化の手法ばかりでなく，デジタル化したコンテンツの発展的な利用も目的としており，そのためにも情報科学における研究結果を応用する事が期待されている．

2.2 自然災害アーカイブ

歴史学や社会学では，自然災害が社会と人間に及ぼす影響について深く研究されてきた．また，多くの研究者によって自然災害に関する年表が作成されてきた．特に地震に関していえば，時系列順に震度，場所，被害状況などの分類を行うことで，地震の検索，地域ごとの地震頻度，地震発生地域の人口との比較を容易に行う事が可能となる．

本学で作成しているデジタルアーカイブの中にも，自然災害に関するアーカイブがある．1つは「史的自然災害に関する資料」であり，もう1つは「史的自然災害に関する資料：掲載資料」である．「史的自然災害に関する資料」に含まれる各コンテンツは以下の組からなる．

ID，和暦，西暦，タイトル，地域，発生時刻，緯度，経度，地震規模，被害状況，備考，関連項目，出典

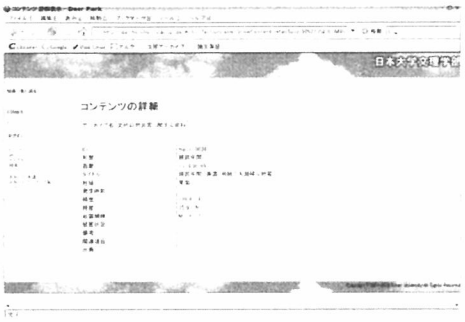

図 1: 史的自然災害に関する資料

アーカイブに含まれるコンテンツの例を図1に示す．ここで，タイトルには地震に関する端的な情報が記述されており，他のインデックスには，タイトルに含まれる地震に対して，出典などから得られる詳細な情報が記述されている．また，「史的自然災害に関する資料：掲載資料」に含まれる各コンテンツは以下の組からなる．

ID，タイトル，掲載資料，原文

アーカイブに含まれるコンテンツの例を図2に示す．図1と同様にタイトルには地震の端的な情報が記述されており，原文には地震に関する記述が原点から抜粋されている．たとえ同じ地震に関する記述であっても出典によっては異なる記述がなされている．すなわち「史的自然災害に関する資料：掲載資料」アーカイブによって原典の比較が可能であり，社会的影響を詳細に考察する事ができる．このように「史的自然災害に関する資料」と「史的自然災害に関する資料：掲載資料」を用いる事で，日時，場所，出典などに関する地震の分類が可能となる．

しかし，これら2つのアーカイブは人手によって作成されており，地震情報の収集には大変労力がかかる．さらに，2つのアーカイブ間にはリンクがはられているが，各アーカイブにおいて災害と災害の関係は述べられていない．例えば，

地震があった (1)

台風の後に地震があった (2)

という記述では大きな違いが予想される．すなわち (2) の場合，社会影響を考える上で，台風

[2] http://da.chs.nihon-u.ac.jp/da/

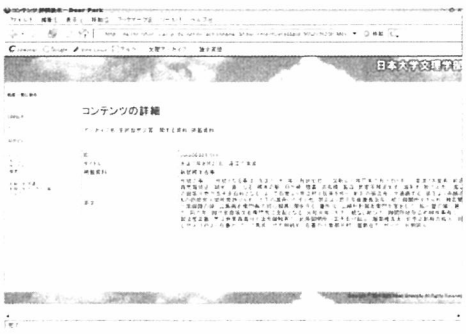

図 2: 史的自然災害に関する資料：掲載資料

という自然災害と地震という自然災害間の時間的な隣接関係は大変重要である．

そこで我々は，自然災害を自動的に収集しアーカイブ化するシステムの構築を目的とする．また，作成したアーカイブに対して時間的な関係を抽出する事も目的とする．ただし，本学で作成中のデジタルアーカイブと区別するために，本稿で作成するアーカイブを自然災害アーカイブと呼ぶ事とする．

3 自然災害アーカイブ作成のための計算機システム

我々は自然災害アーカイブを作成するために，自然言語処理の技術を用いたシステムを提案する．アーカイブの作成に関するシステムは Perl ver.5.8.2 を用いて Vine linux 5.2 CR 上に実装した．

3.1 インデックスの作成

本稿の目的は，事象の時間構造を分析することである．我々は分析対象として新聞記事を用い，特に自然災害として地震に着目した．使用した新聞記事は毎日新聞コーパス[3]である [3]．我々はまず，コーパスから文字列「地震」を検索語として記事と日時を抜き出した．その結果，339489 個の記事のうち，地震を含む記事は 5320 個であった[4]．さらにいくつかのインデックスも同時に抽出する事を試みる．本稿で用いるインデックスを以下のように定義する．

定義 1 (インデックス項目) 自然災害アーカイブは以下のインデックスからなる．

> 記事日時，都道府県，地方，発生日，
> 発生時刻，マグニチュード，震度，
> 震源地，記事本文．

ここで**記事日時**とは新聞記事が発行された日である．**都道府県**と**地方**はそれぞれ地震が発生した場所をさす．特に地方は 9 つの分類 (北海道，東北，関東，中部，近畿，中国，四国，九州，琉球) に分かれている．さらに**発生日**と**発生時刻**は実際に地震が起こった日時をあらわす．

定義 1 に示したインデックス項目はそれぞれの地震を特徴付ける為に重要であり，記事本文に記述があれば同時に抽出する．インデックス項目を自動的に抽出できる事は次のような効果が期待できる．(i) 膨大なデジタルアーカイブに対して，それぞれの特徴を短時間で抽出することが可能となる．(ii) すなわち短時間で特徴を抽出できる事から，インデックスを増やすことで多分野の研究者に対して使用される事が見込まれ利便性が増加する．(iii) 抽出結果を数値的に考察する事が可能となり，必要/不必要なインデックスの選定や，インデックス自体の重要度を考察する事が可能となる．ここでインデックスに対する統語論を定義する．

定義 2 (シグネチャ) 以下の記号とオペレータを定義する．

NI	新聞記事の集合
$ndate(n)$	記事日時オペレータ
$pref(n)$	都道府県オペレータ
$reg(n)$	地方オペレータ
$ddate(n)$	発生日オペレータ
$dtime(n)$	発生時刻オペレータ
$Rich(n)$	マグニチュードオペレータ
$intens(n)$	震度オペレータ
$epic(n)$	震源地オペレータ
$\vee, \wedge$	論理結合子

[3]本研究成果は CD-毎日新聞 98 年版，2000 年版，2001 年版を使用して得られている．

[4]毎日新聞コーパスでは，ページ数ごとに区切られているため，ページ単位でカウントした．

ただし $n \in NI$ である．さらに NI に対して記号 $n_1, n_2, \cdots$ を用いる．また必要に応じて句読点を用いる．

例 1 以下の新聞記事 n_1 が与えられたとする：

> 98'.01.16：１６日午前１０時５８分ごろ、関東地方を中心に地震があった。気象庁によると、震源地は千葉県南部で、震源の深さは約６０キロ、マグニチュードは４・８と推定される。

この時我々のシステムは $\langle ndate(n_1),\ pref(n_1),\ reg(n_1),\ ddate(n_1),\ dtime(n_1),\ Rich(n_1),\ epic(n_1) \rangle$ として，$\langle$ 1998.01.16，千葉県，関東地方，16日，10時58分，4・8，千葉県南部 $\rangle$ を抜き出す．

ここで，組 $\langle pref(n_1),\ reg(n_1),\ ddate(n_1),\ dtime(n_1),\ Rich(n_1),\ intens(n_1),\ epic(n_1) \rangle$ は2.2節で示した「史的自然災害に関する資料」に対応し，組 $\langle ddate(n_1),\ n_1 \rangle$ は「史的自然災害に関する資料：掲載資料」に対応する．

3.2 毎日新聞コーパスの特徴と抽出方法

日本語は他の言語と違い，語と語の間に区切りが無いため計算機を用いて解析するのが困難である．また曖昧性も多い言語である．そこで我々は日本語の形態素解析機 JUMAN [7] と，構文解析機 KNP [8] を利用した．JUMAN を用いる事で日本語の語の並びを形態素，すなわち名詞や動詞といった最小の語の単位に分割する事ができる．さらにKNPを用いる事で，形態素間の係り受け関係を解析する事が可能となる．ただし本稿では複合名詞を１つの名詞とした．また時制を明確にするために，動詞は助動詞付きで１つとした．文字列「地震」と共起する名詞と動詞の例をそれぞれ表１と表２に示す．

表 1:「地震」と共起する名詞の例

7243	地震	989	観測	⋮	⋮
1890	震度	915	発生	28	断水
1555	震源	⋮	⋮	⋮	⋮
1494	こと	176	倒壊	19	地滑り

表 2:「地震」と共起する動詞の例

1055	いう	55	倒壊した	⋮	⋮
857	よると	55	続いた	25	全壊した
799	ある	55	公表した	⋮	⋮
⋮	⋮	⋮	⋮	9	全焼した

ここで，表に含まれる数字は対応する語が毎日新聞コーパスに含まれている数を示している．表１と表２に示すように，共起頻度が高いからといって因果関係が高いとは言えない．そこで時間構造を抽出するために，いくつかのルールを定義する事で解析を行う．詳しくは次節にて示す．

ここではさらに，毎日新聞コーパスの特徴について示す．毎日新聞コーパスはテキストコーパスであり，日付，見出し，記事，形態素解析結果などがタグ付きで含まれている[5]．我々は定義１で示したインデックスを毎日新聞コーパスから抽出する．そのために，定義２で示したオペレータに対して，以下のルールを定義する．

ルール 1 (抽出) オペレータに対して以下の手続きを定義する．

> **n ∈ NI**: 記事本文を抽出する
> **ndate(n)**: コーパスから日付を抽出する
> **pref(n)**: 都道府県と市町村の関係に対するメタデータ (辞書) を別に用意しておく．そのうえで JUMAN と KNP を用いることで，コーパスに都道府県が含まれていれば抽出する．
> **region(n)**: もしコーパスに地方が含まれていれば抽出する．ただし含まれていなくても *pref(n)* があればメタデータ (辞書) を用いて解析する．
> **ddate(n)**: 記事本文に文字列「日」が数字と共に現れていれば抽出する．
> **dtime(n)**: 記事本文に文字列「時」や「分」が数字と共に現れていれば抽出する．
> **Rich(n)**: 記事本文で文字列「マグ

[5] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html

ニチュード」や「M」が数字の前に現れていれば抽出する．

intens(n): 記事本文で文字列「震度」が数字の前に現れていれば抽出する．

epic(n): 記事本文に文字列「震源地は」と「で」が現れていれば，その間の文字列を最小マッチングで抽出する．

システムはルール 1 をコーパスの各ページごとに適用する．

3.3 インデックスによる制約

我々は 3.1 節において，文字列「地震」に着目し，コーパスから新聞記事を抜き出した．しかし抜き出した新聞記事の中には地震の発生とは関係の無いものが含まれる．例を以下に示す．

例 2 地震発生と関係のない記事の例:

> 地震保険めぐる裁判　契約より約款が優先　火災保険訴訟に影響も－－神戸地裁

我々の目的は自然災害の時間構造解析であり，地震の発生と関係無い記事は除外する必要がある．特に，地震を含む複合名詞 (地震調査委員会，地震予知，首都圏直下型地震など) を含む記事の多くは，地震の発生と関係の無い場合が多い．毎日新聞コーパスでは，地震の発生に関する記事のほとんどにマグニチュードもしくは震度が含まれており，場所に関する記述がなされている．よって以下のルールを定義する．

ルール 2 (地方と震度に対する制約)
$\forall n \in NI$, $reg(n) \wedge (Rich(n) \vee intens(n))$ が存在すれば記事を抜き出す．そうでなければ除外する．

ルール 2 によって得られる新聞記事の集合を NI' とする．ここで実験の結果を示す．$reg(n)$, $Rich(n)$, $intens(n)$ が抽出可能な記事の総数は以下の通りであった．

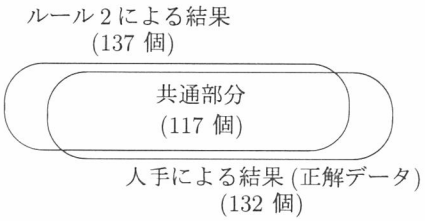

図 3: Recall and Precision

ルール	記事数
総記事数	339489
「地震」を含む記事	5320
$reg(n)$	3439
$Rich(n)$	986
$intens(n)$	1307
$Rich(n) \vee intens(n)$	1780
$reg(n) \wedge (Rich(n) \vee intens(n))$	1531

特に，1998 年 1 月から 1998 年 6 月までの記事に対して，ルール 2 を用いて自然災害アーカイブを作成した．その結果得られた記事数を以下に示す．

総記事数	61637
「地震」を含む記事数	581
ルール 2 の適用後	137

よってルール 2 によって除外された記事は 444 個である．除外された記事のうち，地震の発生記事が含まれていた数は 15 個である．すなわち 96.6%は不要な記事を除外できたが，3.4%は必要な記事も除外してしまった事になる．さらに「地震」を含んだ記事 581 個に対して，人手で地震発生の記事を調べると 132 個であった．ルール 2 の適用結果と人手による結果 (正解データ) との関係を図 3 に示す．

すなわちルール 2 による再現率 (recall) と適合率 (precision) は以下のようになる．

$$\text{再現率} \quad \frac{117}{132} \times 100 = 88.6(\%)$$

$$\text{適合率} \quad \frac{117}{137} \times 100 = 84.1(\%)$$

ここで，再現率とは必要な情報のうち実際に検索された割合であり，適合率とは検索された

情報のうち必要な情報の割合である．一般にこれらはトレードオフの関係があるが，ルール2の適用において正しく判別できなかったものを以下に示す．

例 3 以下の新聞記事が与えられたとする：

98'.04.15　箕面などで地震
98'.04.30　岩手山で、火山性地震

例3に示すように，震度やマグニチュードが書かれず端的な表現がルール2によって除外されているのが分かる．このような記事の多くは，朝刊の記事で端的に地震に触れ，夕刊の記事になって詳細に述べられる事が多い．しかし上記の例についてはこれ以上の詳細な記事が無いため正確な分類が行われない．また，火山性地震や群発地震に関する記事は地震の規模よりも継続性と注意を促す意味合いが強く，地震の規模が書かれる事が少ないためルール2の適用では正しく分類されない．このような問題を解消するには，ルール2を改良し地震記事の抽出条件を考える必要がある．

さらに「記事日時」と「発生日」の関係について注目した．新聞記事の特性上，地震が発生した当日か翌日に記事になる事が多いため，記事内に西暦や月まで記述される事が少ない．そのため抜き出される「発生日」は日付のみが多い．この時，「記事日時」と「発生日」の間には曖昧性が発生する．

例 4 以下の新聞記事が与えられたとする：

00'.01.13：１７日で阪神大震災から丸５年が経過するのを前に、政府は１３日、首都圏直下型の大地震発生を想定した災害訓練を実施した。

例4において，$ndate(n)$ と $ddate(n)$ はそれぞれ13日と17日と抽出される．「記事日時」と「発生日」だけに着目すると，地震の発生が13日より未来の17日に起こるとも考えられる．しかし自然災害に関して言えば，未来における地震の発生が新聞記事に言及されるとは考えにくい．すなわち，このような記事は災害の予定や予測ではなく，過去に起こった自然災害と現在起こる事象との相関によって書かれていると考える必要がある．すなわち「発生日」17日は記事が書かれた13日より過去となる．よって以下のルールを定義する．

ルール 3 (日付による制限)
$\forall n \in NI'$, $ndate(n) < ddate(n)$ であれば n を除外する．

ルール3適用後の新聞記事の集合を NI'' とする．実験として，3年間分の毎日新聞コーパスにルール2を適用した後，ルール3を適用してみた．その結果ルール3によって除外される記事数は23個で，その全てにおいて過去の地震に関する記事，すなわち地震の発生を示す記事ではなかった．我々が作成する自然災害アーカイブには，自然災害が実際に発生した事実のみが重要であり，ルール3によって選別される記事をアーカイブから除外する事で適合率を高める事が期待できる．また例4の場合であれば，「発生日」として13日を抜き出すことも考えられる．この場合は，KNPによる係り受けの関係を参考に $ddate(n)$ の手続きを改良する事が考えられる．

4　自然災害アーカイブの時間構造

3章で得られた自然災害アーカイブは1つの記事に対して1つの組 $\langle ndate(n), pref(n), reg(n), ddate(n), dtime(n), Rich(n), intens(n), epic(n), n\rangle$ を持つ．例えば，火山性地震が連日記事になる場合，新聞のページや発行された日が異なればこれらは全て異なる組として区別されている．すなわち，組と組の間の関係が未定であり，地震が突発的に起きたのか継続的な群発地震なのか判断する事ができない．そこで地震の継続性を判断するために以下のルールを定義する．

ルール 4 (発生日による制限)
$\forall n_1, n_2 \in NI''$, $ndate(n_2) < ndate(n_1)$, $3 < (ndate(n_1) - ndate(n_2))$ であれば n_1 と n_2 は関係がない．

ここで発生日による制限を4日以上としたのは実験の結果得られた数値である．また，$ndate(n)$

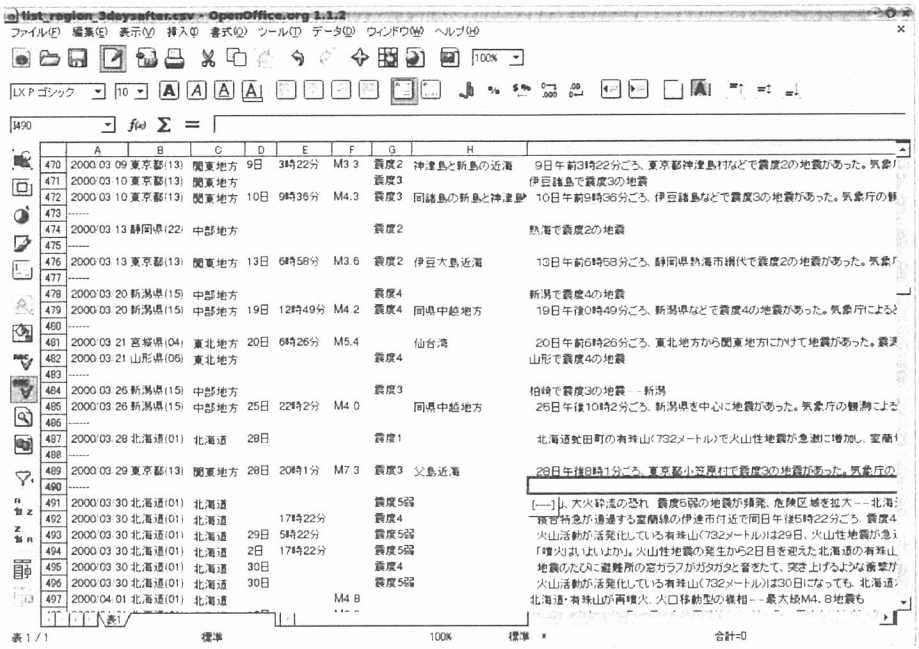

図 4: 時間構造を考慮した自然災害アーカイブ

と $ddate(n)$ が同じ記事であっても $reg(n)$ が異なっていれば異なる地震とするために，以下のルールを定義する．

ルール 5 (地域による制限)
$\forall n_1, n_2 \in NI''$, $reg(n_1) \neq reg(n_2)$ であれば n_1 と n_2 は関係がない．

ここで，ルール 4 とルール 5 を適用した結果を図 4 に示す．図 4 において，最初の列に横線の入った空行はルール 4 もしくはルール 5 の適用結果である．図 4 に示すように，2000.03.30 の記事において北海道の有珠山での火山性地震が同じ区切りの中に含まれている事が分かる．すなわち，これらの記事は一連の地震災害であると考えられる．しかし 2000.03.28 にも有珠山で地震が発生しており同一とする必要があるが，2000.03.29 の小笠原諸島での地震によって地域による区切りが入ってしまう問題がある．そこで地域ごとに自然災害アーカイブを分割して，ルール 4 と以下のルール 6 を適用した．

ルール 6 (都道府県による制限)
$\forall n_1, n_2 \in NI''$, $pref(n_1) \neq pref(n_2)$ であれば n_1 と n_2 は関係がない．

ここで実験の結果を示す．毎日新聞コーパス 3 年間分にルール 4 とルール 6 を適用した結果得られた記事数は以下の通りである．

北海道	42 個	東北	84 個
関東	485 個	中部	169 個
近畿	164 個	中国	109 個
四国	22 個	九州	13 個
琉球	2 個		

さらに，3.3 節と同様の 1998 年 1 月から 6 月までの記事に対しても実験を行った．その結果，次の 7 地方 (北海道，東北，近畿，中国，四国，九州，琉球) においては 100%の分類が可能であった．しかし，関東，中部地方においては正しい分類が行われなかった．その理由としては，(i) 群発地震の多い伊豆諸島，小笠原諸島などが東京に分類され，東京都内の地震と区別できない．(ii) 伊豆群発地震や伊豆東方沖の地震は静岡に分類される，といった理由による．

5 おわりに

我々は新聞記事から自然災害アーカイブを作成するためのインデックスを定義し，抽出実験を行った．その結果インデックスの利用価値の評価が行えた．また，計算機を用いてアーカイブを作成することで膨大な資料を効率的に処理する事が可能であり，様々なインデックスを考察する事でアーカイブ自体の利便性向上が期待できる．加えていくつかのルールを定義する事により，群発地震と突発的な地震の区別を行った．

今後は記事に含まれる事象の言語的特徴やアスペクト的な情報に注目することで再現率や適合率の増加が期待できる．さらに地震以外の自然災害にも着目し，災害間の関係を考察する必要がある．そのためにも情報科学以外の専門家と共に分析を行う必要がある．また，現代語による資料だけでなく旧漢字を含むような資料に対してもアーカイブの作成および利便性向上を試みる．

参考文献

[1] R. B. Allen and J. Schalow. : Metadata and data structure for the historical newspaper digital library, In CIKM '99: *Proceedings of the 8th Intrernational Conference on Information and Knowledge Management*, pp. 147-153, New York, USA, 2005. ACM Press.

[2] J. Barwise and J. Seligman. : *Information Flow*, Cambridge University Press, 1997.

[3] K. Bontcheva, D. Maynard, H. Cunningham and H. Saggion. : Using human language technology for automatic annotation and indexing of digital library content. In *ECDL '02 Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pp. 513-625, London, UK, Springer-Verlang, 2002.

[3] CD-ROM 毎日新聞コーパス 98 年，2000 年，2001 年版：毎日新聞社，日外アソシエーツ

[4] G. Crane and A. Jones. : The challenge of virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection, *Proceeding of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pp. 31-40, 2006.

[5] M. Deegan, E. Steinvel, and E. King. : Digitizing historic newspapers : progress and prospects, *RLG Diginews*, 6(4), August, 2002

[6] M. Federico, D. Giordani and P. Coletti : Development and evaluation of an Italian broadcast news corpus, *Proceeding of the LREC*, v2, pp. 921-924, 2000.

[7] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真, "日本語形態素解析システム JUMAN 使用説明書 version 2.0 ",1994.

[8] S. Kurohashi and M. Nagao : A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures, *Computational Linguistics*, 20 (4), 1994.

[9] M. Markkula and E. Sormunen : End-user searching challenges indexing practices in the digital newspaper photo archive, *Information Retrieval*, 1, pp.259-285, 2000.

[10] S. Yoshioka, K. Kaneiwa and S. Tojo : Occurrence Logic with Temporal Heredity, *Proceeding of the IICAI-03*, pp.1296-1309, 2003

[11] S. Yoshioka, S. Tani, S. Toda : Extraction of temporal relation by the creation of historical natural disaster archive, *Proceeding of the CISE'06*, 2006.

[12] I. H. Witten, K. J. Don, M. Dewsnip and V. Tablan. : Text mining in a digital library. *Int. Journal on Digital Libraries*, 4(1): 56-59, 2004.