知能と複雑系 118−1
(1999. 10. 15)

Intelligent Information Satellites: 予兆発見エージェントの構想と実際

大澤幸生
筑波大学経営システム科学専攻

21 世紀は波乱、そして逆転の時代となろう。その中で未来を予測する情報技術は、重要な国益と平和貢献をもたらす。インターネットのように豊富なデータが分散して蓄えられネットワークで結合されている現状はその機会を与えているが、それを未来予測に活かす上での技術的課題が立ちはだかっている。ここでは、無定型(フォーマット・構造が顕著にばらついている)なデータからなる時系列の中に複数の(原因, 結果)の対があって、それぞれ原因から結果までの時間が大きくばらついている中で、全体として重要な結果に結び付く意味を有する事象を発見することが重要となる。それを実現するための構想が「予兆発見エージェント」である。

Agents Discover Signs of Future: Intelligent Information Satellites

Yukio Ohsawa
GSSM, University of Tsukuba

Prediction of future events is essential for social profits in the chaotic near-future. The distributed, networked, and rich data presented today serves us with chances for such predictions, but new data analysis or mining methods are needed. The available data are unstructured time-series with multiple pairs of cause and effect, each occurring various time length after the other. Our mission is to develop methodologies for discovering significant events for future life of human from these data. Agents proposed here use domain-dependent data processing algorithms for predictions, and refine their results with messages from each other. For prediction of each agent, e.g. for earthquakes, KeyGraph is presented as a reliable algorithm.

1　はじめに

かつて高度経済成長の時代、単調に経済が成長すると言う簡単なモデルに基づく予測ツールがもてはやされた時期があった。それで通用する時代で、世の中は何事もなく進むように見えた... その後エネルギー・環境・経済の破綻が日本を襲い、情報の不完全さがパニックに直結する時代に至った。

ここで情報の不完全さというのは通常の意味で用いている。すなわち、あらゆる状況で意思決定に適用できる情報のことである。状況の変化を未知の因子が決定している場合、その因子が含まれないような情報は不完全であるといえる。いくつか例を挙げてみよう。

スプラトリーにおけるアジアと米国の睨み合いは日本にとって脅威であり、エネルギー確保の技術は日本の命綱となる。

阪神大震災の後ひずみの溜った**花折断層・中央構造線**には、近いうちの震災が発生する可能性が十分考えられる。

市場の**クロックスピード**の変動によってサプライチェーンの見直しをせまられる業界が続出するであろう。日本の場合、変化に対応する新たな技術が急務である。

上の例で太字の言葉を全部知っていた人は意外と少ない。しかし、われわれの祖国日本は最低この3つを含む数多くの要因に強く影響されながら21世紀を進まねばならない。それ故、このような重要な要因が未知ならば発見し、分析してその意味するところを理解することは、危機にそなえチャンスを活かす大きな国益となろう。

そのような予兆を発見する元になるかも知れない膨大なデータが電子化されネットワークで簡単に入手できるのは幸いである。しかし、残念ながら人がその全てのデータを見て、未来を支配することになる未知要因を知るのは極めて時間のかかる作業となる。そこでその作業をどこまで自動化できるか考えよう。現状では、時系列データからの予測技術として単純あるいは純粋なマルコフモデルの他にノイズの各種の考慮を含むモデル、あるいは観測された時系列をそれらの混合のモデルで記述する手法などがある ([広松 1993] など参照)。しかし、これらの従来技術とは異って、われわれに必要なのは、無定型 (フォーマット・構造が顕著にばらついている) なデータからなる時系列の中に複数の (原因, 結果) の対があって、それぞれ原因から結果までの時間が大きくばらついている中で、全体として重要な結果に結び付く意味を有する事象を発見することである。それを実現するための構想が「予兆発見エージェント」である。

2　予兆発見エージェント

今あなたが酒屋の店主で、「この冬、どんな酒が売れ筋だろうか」と悩んでいるとしよう。この読みを外すと在庫を抱え、自分が年を越せなくなる。現在シェアを延ばし続けている銘柄が近い未来にも売れ筋として生き残る可能性は高いかも知れない。しかし、この冬がもし暖冬だったら例年よりビールが多く売れる一方、熱燗むきの日本酒にはあまり期待できない。景気がどんどんよくなってくれたら高級なブランデーも少しは売れるだろうし、日朝関係がよくなれば平壌焼酎も旨いから流行るかも知れない...

このように、酒の売れ筋予測だけにもブーム・気象・経済・国際政治といった複数の要因を考慮しなければならない。逆に、酒が売れるというような一見小さな事象が感情の交際を経て国際交流の起点になることも有り得る。すなわち、さまざまな事象の予測は互いに因果関係を有しており、中には双方向の因果関係もあるということである。

このように複雑な未来の予測を、マルチエージェントの枠組みで実現しようとするのが「予

兆発見エージェント」である。一匹のエージェントは様々なデータ(インターネット上に公開されたデータだけでも膨大である)を渡り歩き，自分の得意なデータから発見できる予兆を見出す．自分(例えば新聞記事データ担当エージェント)にはわからないデータ(地震データ)には，それ専門のエージェントを呼びつける．そういうエージェントの見出した重要な予兆を総合的に判断するのは人間だが，エージェント間で発見成果を交換した上での結果をその人間に差し出すことは判断の強力な支援となろう。

ここで言う発見は「これまであまり起こらなかったがこれから起こる可能性があり，人間にとって重大なできごとの予兆を見出す」ことであって，「これまで起こっていたことのパターンを学び，同様のことが起こる予想を下す」ためのいわゆる学習とは異なっている．後者は人工知能における機械学習や統計解析として研究されてきたが，発見エージェントは前者を行なうべきエージェントである．

2.1 予兆発見エージェントの構成

有限・複数の予兆発見エージェントがエージェンシーを構成する。現状で考えているのは以下の5つである。

気象予測エージェント クエリーに対し、民間気象予測企業のデータから該当する期間の予測データ Y_1 を戻す。

地震予測エージェント 地震時系列データから、近い未来に危険となる活断層 Y_2 を見積もって選択する。

売れ筋予測エージェント POSデータ X_3 と、考慮する期間 T に対する Y_1 の関係を見出す。この関係と現在の気象予測 Y_1、Y_4 (ブームの語は宣伝媒体に乗せやすいので) から売れ筋商品 Y_3 を予測する。

ブーム(宣伝効果)予測エージェント Y_1, Y_2, Y とWebページ(特に広告など注目度の高い場所に出現する語に重みづけする)の出現語データ X_4 から、近い期間におけるブームとなる語 Y_4 を予測する。

景気予測エージェント 大変複合的なので、現状では入力と出力にあたる変数の見通しが立っていない。本来はこの出力変数は売れ筋予測とブーム予測に影響すると考えられるし、Y_2 はここでの入力変数の一つとなる。

データの受渡しが入り組んでいるが、その具体的な方法は各エージェントが他エージェントからの情報をどのように用いるかに依存する。例えば、大きな地震が東京で起きるのと京都で起きるのとでは、その後の大衆の関心は変って来る。いずれも人命の尊さや防災のあり方を再考させる点は同じであるが、東京の大地震は政治・経済の面で世界を揺るがし物流れを停滞させる度合が大きい。その売れ筋やブームへの影響は計り知れないが、当面をシンプルなモデルでデータを売れ筋予測とブーム予測エージェントに受渡し、その後学習を経て精緻化しなければならない。

例えば、総合的に各種日用品や嗜好品まで売るデパートやコンビニエンスストアの売れ筋予測エージェントに地震エージェントが与える情報を述べてみる。地震が地方都市Aで発生する可能性が高いとすると、都市Aの店舗は嗜好品・贅沢品の発注を抑えるためサプライチェーンの見直しが迫られる。また、都市Aに発注先を持つ店舗は別の発注先を開拓しておくのも必要な戦略となる。一方、地震が東京のような影響力の強い都市で発生するのならばそれではすまず、全国的に経済的打撃を与えるから、全国の店舗で販売を必需品に押さえる必要が出てくる。そこで、東京で地震が起きると予測するならば、地震予測エージェントはここに述べた簡単な因果ルールを用いて作成した「全ての店舗で嗜好品・贅沢品の発注を押さえ、各店舗は東京以外の発注先を開拓せよ」なるメッセージを売れ筋

予測エージェント（当面はデパート・コンビニエンスストア・スーパーマーケットに限る方針）に伝える。式（??）は地震と直接因果関係のあるものだけに関するメッセージだが、この他に間接的な因果関係の影響も発見的に追加していく。売れ筋エージェントは、他のエージェントからのこのようなメッセージと自分のPOSデータ解析結果を統合して結果を出力する。

このように、各予兆発見エージェントは自分自身のデータ解析をメッセージ交信によって予測結果を更新していく。ところで、景気予測エージェントは当面見送ることにしても、なおかつ現状では（単独のデータ解析で）実現手法の見とおしが立っていないのが「地震予測」と「ブーム予測」である。これに対して現在、以下の手法に述べるKeyGraphの適用について研究を進めている。

3 KeyGraph:「主張」を抽出するインデキシング

まず、KeyGraphなるキーワード抽出アルゴリズムを紹介する（一見前節と繋がりが悪いが、とにかく先を読んで頂きたい）．元々，KeyGraphが狙うキーワードとは文章の主張を表す単語である．文章というのはその著者が言葉で考えを表現していく一種の時系列であり，その文章の中で主張とは何度も現れないが新たに打ち出される重要な概念である．それは丁度，事象の時系列の中でわずかに出始めた，しかし後で重要な変化の元になるような現象，つまり「予兆」に似ている．例えば，先述の「スプラトリー」はこれまでインターネットにも殆んど出て来なかった言葉であるが21世紀の世界を決めるキーワードであり，最近の新聞記事では国際欄を急に賑わせ始めている．

また，過去にほとんど揺れていないが少しずつ周囲の地殻に押されて揺れ始めている活断層は大地震をひき起こす可能性が高い．KeyGraphとはただ文章からキーワードを取り出すだけではなく，過去一年の新聞記事から「スプラトリー」を，過去数年の地震データから未来の危険断層を発見し，劇的変化の予兆を人に知らせる可能性を秘めている．

筆者は、このKeyGraphの動作を「ブーム予測」「地震予測」エージェントの中で、他エージェントと独立にデータ解析する部分のアルゴリズム（これがシステムの駆動力である）一つと考えている．KeyGraphを、ここではまず当初提出したとおり文章からキーワードを抽出する手法として説明する。

3.1 KeyGraphによるキーワード抽出

文章を建物に喩えるとKeyGraphは

> 建物が立つには，**土台**（文章が基にしている基本概念）が必要である．壁（文章の構成に必要な説明部分），ドアや窓（詳細な記述），様々な装飾（比喩や例など，付加的な記述）もある．しかし，建物の本質は日射や雨から住人を守る**屋根**（主張点）であって，屋根を支えるために柱（内容の主な展開）がある．

という仮説に基づく．文章の中で繰り返される頻出語には文章の主張として筆者が用いた単語も含まれるが，それ以外にも文章の主張を支えるための重要な概念として文章の「土台」を形成する単語が数多く含まれている．これら土台の上に立つ「柱」に支えられて文章全体の論点となっているのが主張（「屋根」）である．ここで多くの「土台」に支えられている「主張」を表すキーワードを抽出するために，土台と主張の間の関係である「柱」をもとにしたキーワード抽出法がKeyGraphである．KeyGraphの手続きは次の3フェーズからなる．

1) **土台の形成：** 文章Dの形成の準備あるい

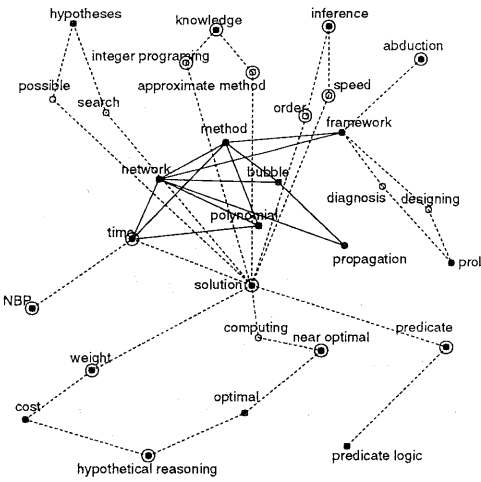

図 1: KeyGraph の出力結果

は前提となる基本概念（具体的には，後述の語の共起グラフにおいて強く連結しあう語の集まり）を**土台**とする．

2) 屋根の形成： 1) で取り出した土台たちに強い (柱の) 力で支えられて文章を統合する語を**屋根**とする．

3) キーワードの抽出： 土台と屋根を結ぶ強い柱が多く集まった語を D のキーワードとする．

これらの各フェーズの詳細については [大澤 99] を参照されたい。図 1に KeyGraph を用いてキーワード抽出を行なった結果を示す．

図中の二重丸がキーワード，黒丸が土台の語で白丸が屋根，実線が土台中のリンク，点線が強い柱を表している．文章 D が述語論理仮説推論を整数計画問題に帰着して計算速度を改善し，多項式オーダー時間で解く近似解法についての論文であったので，
predicate(述語)， time(時間) などの単語に加えて，出現頻度の低い
speed(速度)， approximate method(近似解法)
などがキーワードとして取り出されていることは，主張を正しく抽出する KeyGraph の性能を表している．KeyGraph によるキーワードを用いてキーワード検索サーチエンジンを作成し、定量的に性能評価を行った結果も??にある。

インターネット上の文書が分類しにくいといわれる原因の一つは，個性の主張の場としてもインターネットが活用されるほどに普及しているからである。この意味では、KeyGraph のように著者の個性・主張を重視した文献検索，文章の理解の役割はこれからも大きくなっていくであろう．しかし，単語が書かれた時系列データとして存在するテキスト情報の構造に着目して要所を切り出す方法の効力は文章からのキーワード抽出に留まらない．さまざまな分野への応用の可能性を秘めている KeyGraph の応用例を次に紹介する．

4 時系列データからの発見ツールとして

歴史に関する文書集合から，KeyGraph によって歴史文書の中に生じている「土台（原因）→主張（結果）」という因果関係を取り出し、一つの文書の結果が続く文書においては原因になるという特性に着目することで歴史のつながりを見出して文書を並び替え，歴史ストーリーを生成するシステム HiStory[大澤 98a] をこれまでに作成した。これは、過去について書かれた文書から、その文書中の一連の事象の結果と位置付けられる事象を見出したものであった。

ならば、未来に対してはどうか。実は、過去のデータを見て、未来に重要となる要素を見出すといういくつかの用途にも KeyGraph は有効さが示されている。たとえば、デスクトップで雑多なファイルを使うユーザに、重要なファイルとその間の関係を見やすく表示し、近い未来に重要となるファイルも見せるシステムを構築した [大澤 98b]。これは、文

章がいくつかの根拠となる文の集まりに支えられて著者の主張する考えが述べられているのと同様に，時を接して用いられることの多いファイルの集まり (実験用プログラムなど) はユーザの作業の土台となっているという仮説から、ユーザのファイル使用履歴を KeyGraph に入力して，ユーザにとって必要なファイルとそれらファイル間の関連性を抜き出すものである。

筆者はこれに似たアナロジーの仮説から、さまざまなユーザの WWW ブラウジング歴から本質的な興味の全体を見出し、その総和を多数のユーザについて獲得することによって大衆の興味の本質の動きをとらえるのを「ブム予測」のデータ解析部のアルゴリズムと捉えてしている。

4.1 地震履歴から危険断層の発見

地震の履歴から今後の地震を推測する方法として，地震の空白域 (最近大きな地震が起きていないが，その周辺では起きているような地域あるいは海域) を危険地域とみなす方法がある [Ohtake 1994]. しかし，その領域には本当に震源となる理由がないのかも知れない。東海地震は実際，発生が予言されてから既に 20 年以上が経過しているがまだ発生していない（もちろん、これから起きるかも知れないが）.

地震履歴を入力データとして確率的な推定方法を適用した例も多い [Rikitake 1976] が，比較的限定された地域での履歴から同じ地域の危険度を予測するものが多い. ところが実際には，地表近くに現われていない伏在断層や地殻の歪みを介して断層の動き同士は関連し合っている. そもそも断層地震というのは，それら断層近辺の力の関係によって発生するのである.

筆者の開発した方法は、このような相互作用を KeyGraph でリンクとして得ることにより、近い未来の地震の発生する危険の高い

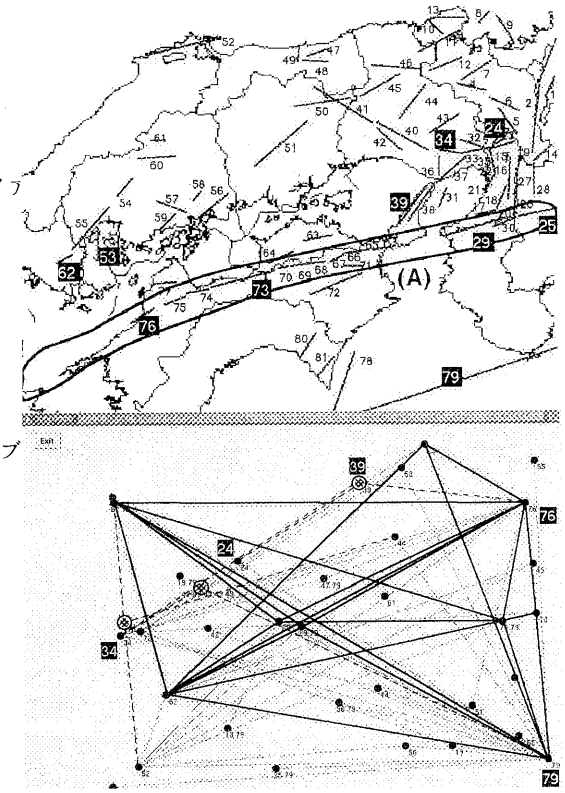

図 2: F^3 の関西地方に関する結果

断層を「主張」として求めるものである。

4.1.1 KeyGraph による断層危険度評価

KeyGraph による断層危険度評価システムを Fatal Fault Finder (F^3) と呼んでいる. 概して，大きな歪みが蓄積されて今まさに動き始めつつある活断層が近い未来の大地震の震源となる可能性が高いので，F^3 もそのような断層を見出す仕組みを KeyGraph で具体化している.

F^3 は先ず，地震関連の観測所から提供される地震履歴データの各地震で活動したとみられる活断層を地震発生順に羅列する. すると，地震履歴は単語列となる（断層名が各単

語となる）．ここで重要なことは，比較的規模の大きな地震は地殻活動の区切れになると考えられることである．というのは，大規模な地震によって大きなエネルギーが放出され，ある部分の地殻の移動の速さと向きを変化させるからである．そこで一定のしきい値以上のマグニチュードの地震発生の箇所にピリオド（.）を挿入すると，そこにできる文字列は形式的にだけでなく，地震の履歴にとって意味の在る「文章」となる．

すなわち，人の書く文章では考えの基礎となるいくつかの土台の上に主張が立てられるのと似て，地震履歴では普段から動いているいくつかの地域の地殻に挟まれた断層が圧迫されて新たな大地震を起こすと考えられる．更に文章の土台は同時出現する単語からなるというKeyGraphの仮定は，ある限定された時期に発生する地震が共通の原因である地殻運動から発したものであるという地震についての仮定と対応する．それ故，地震履歴から得られた文章をそのままKeyGraphで処理すると，そこで得られるキーワードは新たに大きな地震の発生する可能性のある活断層を示すのではないかという期待をもって実行した．その結果が以下のとおりである．

1985年から1992年までの地震履歴データについて上記の処理を試みたところ，図2，3のような結果を得ている．数字のついた各ノードが活断層で，黒いノードと実線リンクがKeyGraphの「土台」を構成する．点線が「柱」で二重丸ノードが「屋根」すなわち要注意断層である．

図2の上図は関西地方の地図で数字をつけた太い実線が活断層である．下図でF^3が危険とみなしたNo.39は阪神淡路大震災の震源となった野島断層で，この図はNo.39がNo.24やNo.34の北側の断層の動きとNo.76やNo.79といった南側の断層の動きの間でストレスを受けて少しずつ動いていることを示している．

図3は全国でF^3が危険とした断層の位置である．各枠内で起きた地震のデータから計算したのが実線，日本全体の地震から計算した結果が破線である．1993年の北海道南西沖地震（奥尻島大津波）や1995年の兵庫県南部地震（阪神淡路大震災）の震源となった野島断層が高い重要度を持つキーワードとして抽出されている．さらに，今後の動向が要注意とされる四国の中央構造線などが抽出された．実線と破線とはそれぞれ，小さな地域の運動の間に挟まれた活断層（領域A，Bなど）による地震と，大きな地域の運動に挟まれた活断層（海域Eなど）による地震を表しているものと見られる．この区別は，実線よりも破線の方が危険が大きいということとは意味が違うことに注意したい．

5 結論

わが国の国益をつかさどる柱としての政治・経済・地球環境における危機管理と機会活用のために、未来を総合的な観点から予測するマルチエージェントの枠組みを提案し、地震予測とブーム予測の初期成果をまとめた。様々なデータにそれぞれふさわしい発見エージェントを研究者が手分けして作り，発見エージェントたちがどうやって総合判断を形成するかを共に考察し社会貢献を達成することで、国際貢献できる国力を培うことができれば幸いである．

参考文献

[大澤 98a] 大澤幸生, 村上尚央, 谷内田正彦：内容における因果関係を用いた文書集合からのストーリー抽出, 第33回人工知能学会SIG-FAI研究会資料, pp.43 – 48, (1998).

[大澤 98b] 大澤幸生, 須川敦史, 谷内田正彦：グラフに基づくキーワード抽出法KeyGraphのデスクトップ整理への転用, 第112回情報処理学会SIG-ICA研究会資料, (1998).

[大澤 99] 大澤幸生, ネルスベンソン, 谷内田正彦: KeyGraph: 単語共起グラフの分割・統合によるキーワード抽出, 電子通信学会誌論文誌 J82-D1, No.2, pp. 391 – 400, (1999).

[Ohtake 1994] Ohtake, M.: Seismic gap and long-term prediction of large interplate

earthquakes, *Proc. of International. Conf. on Earthquake Prediction and Hazard Mitigation Technology*, pp. 61 – 69, (1994).

[Rikitake 1976] Rikitake, T.: Recurrence of great earthquakes at subduction zones, *Tectonophysics*, 35: pp.335 – 362, (1976).

[広松 1993] 広松 毅・浪花貞夫「経済時系列分析の基礎と実際」多賀出版 (1993)

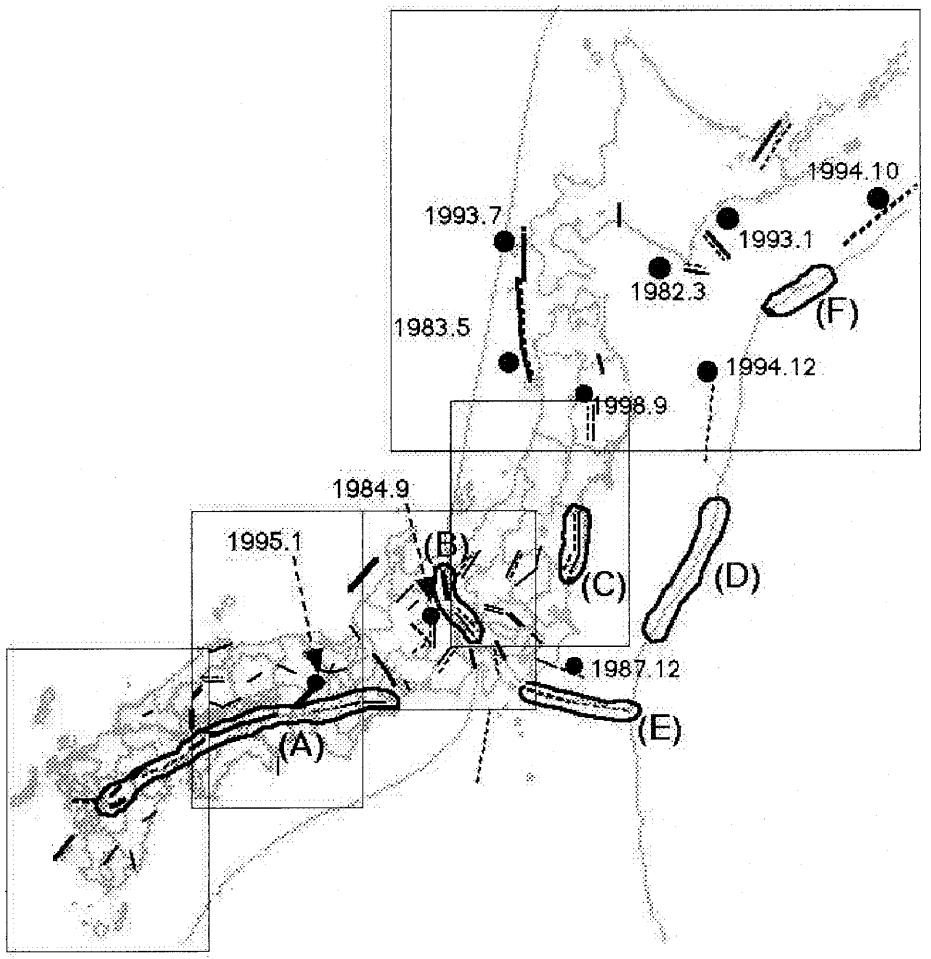

図 3: F^3 が危険と判定した断層の位置