知識蓄積型推論データベースの開発、応用

The outline of a knowledge accumulation database; Synapse engine

近 藤 克 彦
Katsuhiko KONDOU

株式会社タクミ　代表取締役（〒104-0033 東京都中央区新川二丁目９番９号　SH ビル６Ｆ
E-mail: kkfunk@tacmi.co.jp）

ABSTRACT. This database is a search engine that performs management and reference focusing on relation of language and language. It seems that the simulation of being reminded of another language from language with a man is carried out. The conventional database was performing classification and search in the restricted category. In this case, when a new category occurs, change of the database structure itself is needed. With synapse engine, there is no concept primarily called category and the combination of two languages can be registered without any restriction. Moreover, since it is manageable how many times a specific combination was registered, the strength of combination can be seen. Moreover, since it also has the time registered at the end, it is possible to give priority to a new combination. It is the simulations of associating one after another according to the strength of relation between thinking, just like a man thinks of something and the following idea is born from the idea.

1． 背景

インターネットを柱とする情報化社会となってきた現在、既存のデータベースは進化の過程において勘定系からの発展をしているものが主流をなしている。インターネットは情報ハイウェイ構想の上で成り立っており、その生業はまさにWEB化されている。
この環境の中で現在用いられているインターネットの検索エンジンはあくまでもサイトの登録者によってカテゴライズされており、サーチをする使用者側からのアプローチによる検索形態になっていない。これは大量のデータベースを扱う場合既存のデータベースの登録がツリー構造を主とした検索型になっているために検索用の登録を行う必要があることに起因しているケースが多い。また、マルチメディア要素、XMLなどの最近主流になりつつあるデータ形式に対しての検索も弱い。
このような背景の中で求められる検索は、
①高速な検索であること。
②関係式により検索ができること
③あらゆる形式のファイルに対応できること
これらの求められている検索を実現するためには既存のデータベースにおいては構造的に不可能に近いと思われる。
これらの機能を実現するためには新しい構造を構築する必要がある。

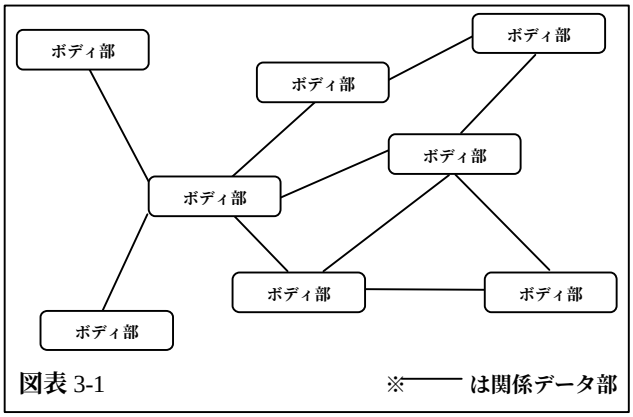
2． 目的

背景を省みたときに新たなる検索エンジンとデータベースの必要性があり、既存値のデータベースとは異なった構造を持つ必要があると考える。 特に過去のデータベースに見られる検索系やデータ構造は現状必要とされているマルチメディアやインターネットで使用されているデータに対しては非常に使い勝手の悪いものとなっている。
また、インターネットのようなWEB構造を持つようなデータベースを現状では一般的に見ることはない。そのためWEB構造をもつようなデータベースを作成する必要があり、かつデータベース内を快適に検索できるような検索システムを構築することが重要となる。
これらの事象よりインターネットの環境においてB to BやC to Cのビジネスに即したデータベースを作成すること、およびインターネットの特徴であるリンクを主体としたデータベース検索システムを構築することを目的とする。

3． 提案内容

本データベースは、関係式を主軸にデータの繋がりをシノプシス系に類似した2次元マトリクスの形式で構成される。データ構造としては関係式を非常にシンプルな構造で表現している関係データ部と写真、動画、テキスト、XMLなどの各種形式のファイルに対応したボディ部から成り立つ。
これらの関係式は下図3.1に表記しているようなシノプシスの構造に類似した2次元のマトリクス構造の上でデータ検索が行われる。検索の主体はボディ部からの関係式の検索と、関係データ部からのボディ部の検索の2種類の方法が考えられる。
これら2種類の検索からはボディ部とボディ部の繋が

りの強さや各ボディ部に対する繋がりがどのように行われているかを検索することが可能となる。また関係データ部からの検索は同一の関係を持つボディ部がどのくらい存在しているか、繋がりの強さが各ボディ部に対してどの程度存在しているかを検索することが出来る。



図表 3-1　　※――― は関係データ部

4.　成果物

本年度は全体システムにおける基本的な動作部分の完成を目指す。これは、データベース検索において高速化するためのアルゴリズムの調整を含む。検索においては関係式に主眼をおいての開発とする。

・高速検索用エンジン　（プロトタイプ版）
・知識蓄積型推論データベース(シノプシス型データベースエンジン）（プロトタイプ版）
・インターネット検索用検索エンジンプログラム（プロトタイプ版）

5.　動作環境

環境として考えられものは、一般的な LAN 環境やインターネットの環境が望ましい、基本的なスペックとしては以下のようなことが必要とされる。

・Windows95,98,2000,UNIX などの OS 上での動作
・AT 互換機、SetTopBox 上での動作
・TCP/IP プロトコル上のネットワーク

6.　重要ポイント

本データベースにおいての重要ポイントを、下記に示す。

① 関係式の強さおよび繋がりを主軸としていかに高速に検索を行うこと。
② 多数のデータの関係式を蓄積型に溜めることにより人間の持つ知識や経験値をデータベース化すること。
③ インターネットにおける WEB 構造をデータベース化することにより既存データベースでは行うことの出来なかった、関連、関係による分析を行うことが出来ること。

この３点が最も重要なポイントとなる。
①においては、基本的な高速化のためのアルゴリズムの開発、およびチューニングを行う必要がある。
②においては関係式と言う既存のデータベースにおいて実現していないインデックスを構造体として持つ必要がある。
③おいては、関係式を主体とした②の構造の内部においての検索および分析を行う必要がある。

7.　シナプス型データベースエンジンについて

(1) ソフトウェアの概要

このデータベースは言葉と言葉の「繋がり」に注目して管理、検索を行うエンジンである。人間がある言葉から別の言葉を連想することをシミュレーションしているとも言える。従来のデータベースでは、ある言葉に対していくつかの制限されたカテゴリ分けを行うことが可能であったが、その場合新規のカテゴリが発生した場合にデータベース構造自体の変更が必要になってくる。シナプスエンジンではそもそもカテゴリと言う概念が無く、2 つの言葉の組み合わせを無制限に登録することが出来、さらに特定の組み合わせが何度登録されたかも管理出来るので、組み合わせの強さを見ることが出来る。また、最後に登録された日時も持つので、新しい組み合わせを優先することなども可能である。
これにより、人間が何かを思いつき、その思いつきから次の思いつきを、つながりの強さに応じて次々と連想する事をシミュレーションすることが出来る。

(2) データベースエンジンの構造

データベースエンジンの構造は以下のようなフローを

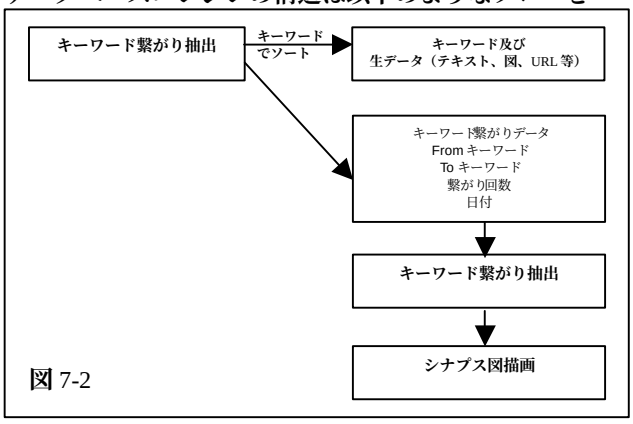


図 7-2

もとに構築される。

(3) データの登録方法

現在はテキストファイルに直接入力するか、元になる CSV 形式のファイルからコンバータツールを使って登録。さらにあいまい検索用のインデックス作成を別途行う。最終形としてエントリ画面を用意することは簡単なので、それに合わせて書き直すことも可能。

(4) データの構造

a)キーワード、テキスト

・キーワード　　可変長文字列（例えば見出し語句）
・テキスト　　可変長文字列（実際のテキスト。例えば語句に関する解説）

・URL　　可変長文字列（例えばホームページ URL や、ワープロの文書ファイル名など）但し、現在は実装されていない。

b)From キーワード繋がり

・From キーワード　　可変長文字列（次項のキーワードへ繋がるキーワード）

・キーワード　　可変長文字列

・繋がり数　　同じ From キーワード、キーワードでの繋がり回数。

・日付　　最終登録日

c)To キーワード繋がり

To キーワード　　可変長文字列（次項のキーワードから繋がるキーワード）

キーワード　　可変長文字列

繋がり数　　同じ To キーワード、キーワードでの繋がり回数。

日付　　最終登録日

(5) アプリケーション構造

アプリケーションの構造は以下のようなフローをもとに構築される。

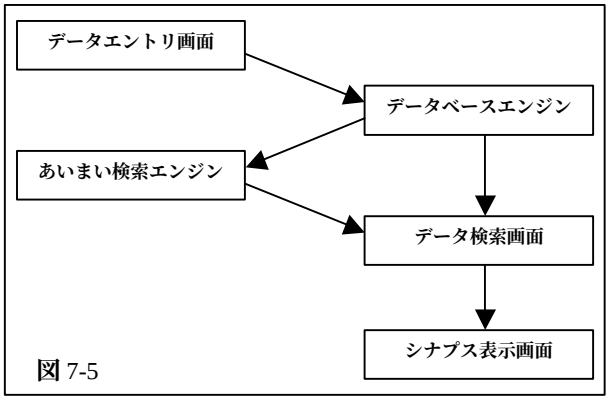
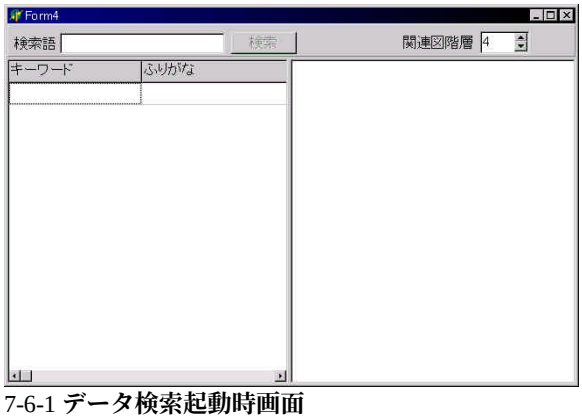


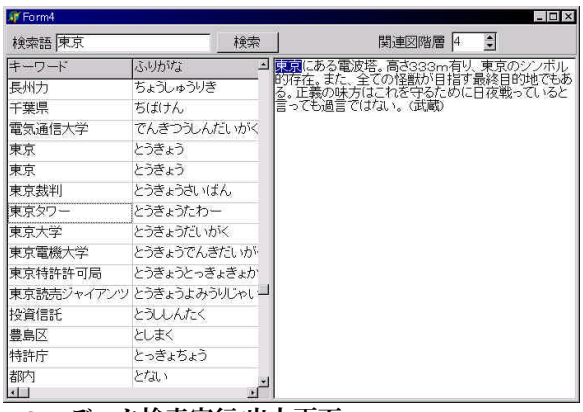
図 7-5

(6) アプリケーション画面例

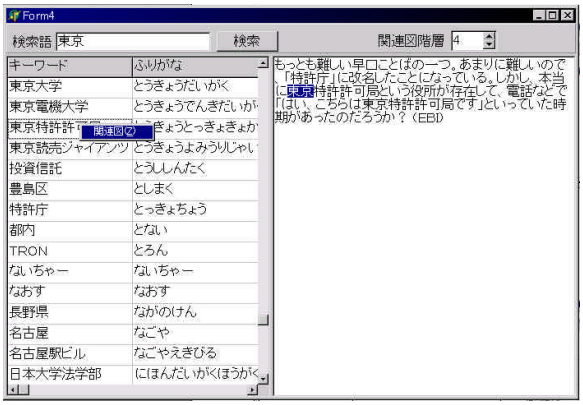
シナプス型データベースエンジンのアプリケーション画面の例として以下のような表示が例としてあげられる。



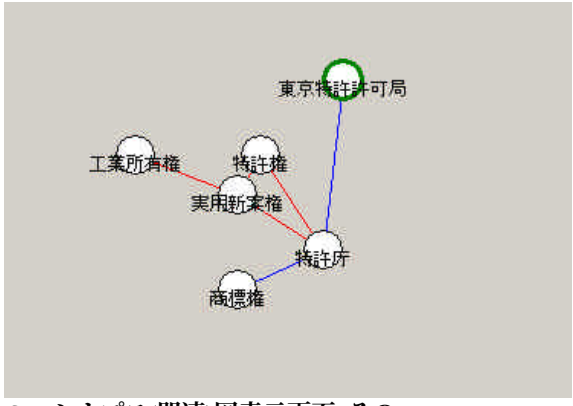
7-6-1 データ検索起動時画面



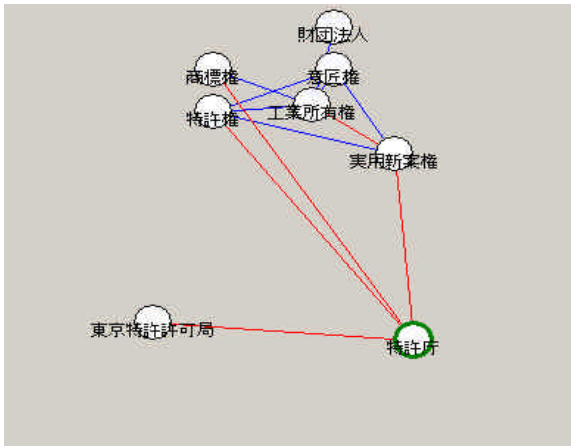
7-6-2 データ検索実行/出力画面



7-6-3 シナプス(関連)図表示起動画面

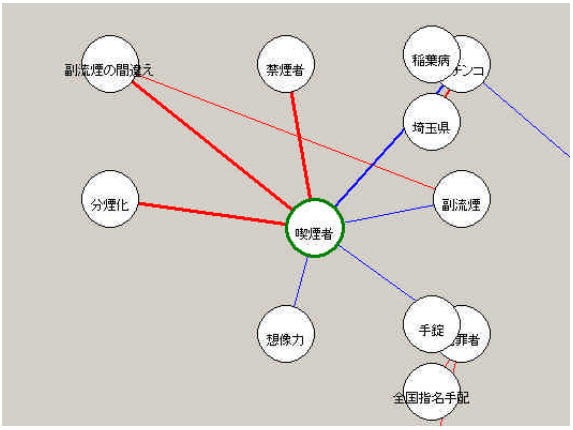


7-6-4 シナプス(関連)図表示画面 その 1
“東京特許許可局”をキーワードとして表示

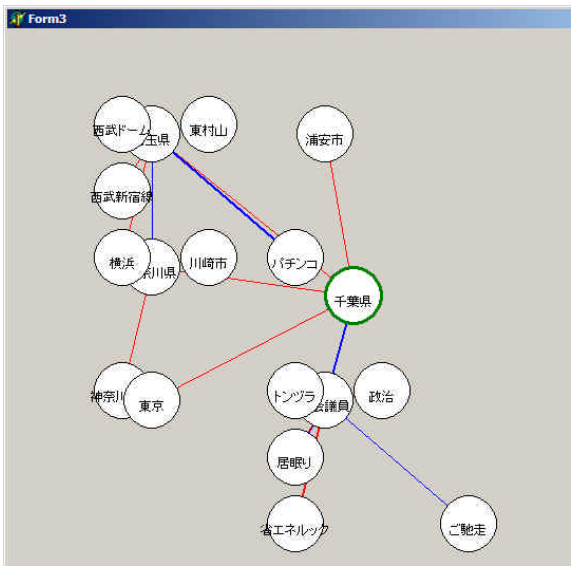


7-6-5 シナプス(関連)図表示画面 その 2

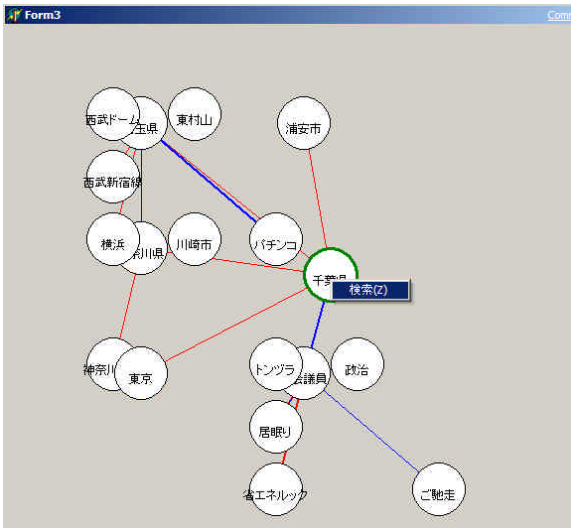
“特許庁”をダブルクリックすることによりキーワードを切り替えて表示



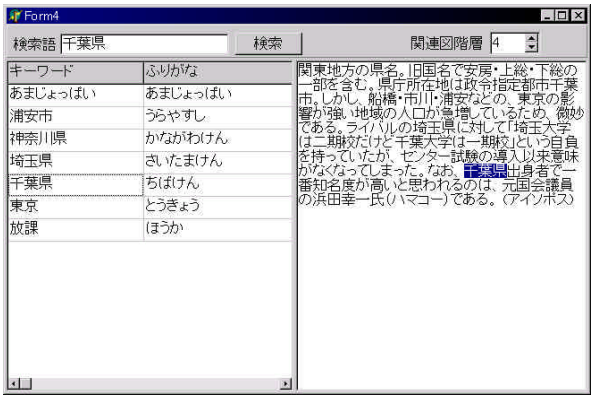
7-6-6 シナプス(関連)図表示画面 その 3
検索経路が蓄積され、繋がりの強さが線の太さで表示される



7-6-7 シナプス(関連)図表示画面 その 4
検索経路の距離がキーワード間の距離で表現されている



7-6-8(1)



7-6-8(2)シナプス(関連)図表示画面 その 5
シナプス図上のキーワードから検索を行うことも可能

(7) 用途例

・国会議員　国民により選挙され、国会を構成する議員。衆議院議員・・・
・衆議院議員　衆議院を組織する議員。代議士とも称する。任期は四年・・・
・代議士　国民から選ばれ、国民の意見を代表して国政を議する人・・・
・国政調査権　憲法上、国会の各議院が有する、国政に関して自ら調査・・・
・日本国憲法　第二次大戦の敗戦後、大日本帝国憲法を全面的に改正し・・・
・内閣総理大臣　内閣の首長たる国務大臣。国会議員の中から国会の議決で指・・・

上記のような辞書データがあり（実際には本文はもっと長い）、これを検索して意味を調べるような場合を考える。
“国会議員”を検索してみると、“内閣総理大臣”、“国会議員”が検索される。“内閣総理大臣”の本文を参照することで、“国会議員”→“内閣総理大臣”の関連が作られる。“国会議員”→“国会議員”の場合、同じ語句なので関連は作成されない。
“国会議員”の本文を見ると、なかに“衆議院議員”の語句があるので、これを新たに検索すると、“代議士”、“衆議院議員”、“国会議員”が検索され、同時に“国会議員”→“衆議院議員”の関連が登録される。
同様に、“代議士”、“国会議員”の本文を参照することで、“衆議院議員”→“代議士”と“衆議院議員”→“国会議員”の関連登録が行われる。さらに“国会”で検索し“内閣総理大臣”、“日本国憲法”、“国政調査権”、“国会議員”などが検索され、関連が登録される。
この時点で“衆議院議員”をシナプス図で表現すると、下記のように表現される。

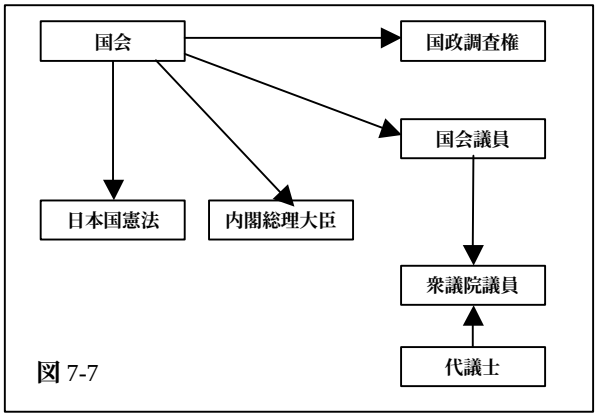


図 7-7

このとき、例えば“国会議員”→“衆議院議員”の関連が何度も登録されると、シナプス図では接続線が太く表示され、つながりの強さを表す。

8． 結論

本シノプシス型データベースはマトリクス構造を持ち、関係式を主体としたデータの繋ぎこみを行うことにより過去のデータベースでは実現することの出来なかった多次元的なマトリクスによる繋ぎこみの強さによる検索を行うことが出来る。
具体的な例で述べると過去のインターネット検索エンジンでは、キーワードによるコンテンツ上でのワード検索が主であったが、本データベースにおいては同じキーワードを用いたとしてもキーワードの使われ方により次にどこに繋いでいるのか、またはどこからの情報によりこのキーワードに繋がってきたのかなどの関係式を主体とした検索を行うことが可能となる。また、同一のキーワードからの繋ぎこみの頻度が上がるほどデータベース自体が学習を行い、強い関連を持っている先を視差するようになってくる。これは、インターネット社会において自分が興味を持って調査を行った内容から他人がどう繋いでいるかを知ることにより、関連する周囲の情報を関係式の強い順に知ること出来ることとなる。この推論はシステムがファジー論理やAIなどの論理で行うシステムの推論ではなく人間が持っている経験値を他人が共有することになる。
このデータベースを用いてデータの蓄積を行うことにより、人間の経験値や既存の知識にのっとった繋ぎこみを行うことができる。これらのことによりインターネット社会における人間の知識や経験を踏まえた上での新しいマーケットを切り開くことが出来るようになる。また、職業としても現在サーチャーと言う職業に対してカテゴライザーといういわゆる経験値や知識の豊富な人のデータの繋がりを有償で行うことにより、新しい職業の開発と新しいマーケットを構築することが可能となる。

9． 参加企業及び機関

株式会社タクミ

10．参考文献

無し。