自 然 言 語 処 理　129－1
（1999．1．20）

ConceptBaseの言語処理と新しいソリューション

野村直之
(株)ジャストシステム
〒107-8640 東京都港区北青山 1-2-3 青山ビルヂング
03-5412-3993　fax:03-5412-3988
Naoyuki_Nomura@justsystem.co.jp
http://www.justsystem.co.jp/cb/index.html

概要： ConceptBase はヴェクトル空間法と転値ファイルによる類似文書検索をコアにもつシステム。高速性と高精度を達成するために、複合語句（Concept)間の部分マッチング、関連語抽出の近似処理、などの 独自の工夫を施している。その概要とともに、対象とする文書空間のスケール拡大のための新しいソリューションとして、自動分類機能と、複数文書の鳥瞰ビューを提供する最新の自動要約機能を紹介する。

ConceptBase - A NL-based IT Solution Core

Naoyuki Nomura
Justsystem Corporation
Aoyama Bldg. 1-2-3 Kita-Aoyama Minato-ku Tokyo 107-8640 Japan
+81-3-5412-3993　fax:+81-3-5412-3988
Naoyuki_Nomura@justsystem.co.jp
http://www.justsystem.co.jp/cb/index.html

Abstract: ConceptBase is an IT solution core based on VSM（Vector Space Model）and inverted file indexing. The performance has been improved by concept matching, which takes advantage of analyzing and comparing the structure of complex noun phrases. This paper shows how natural lauguage analysis technologies enhance the accuracy of concept search. Also introduced are a couple of new applications of ConceptBase - CB Classifier and CB Summarizer.

1 はじめに

近年、WWW(World Wide Web)の技術標準に基づくネットワーク利用の拡大にともない、伝統的な情報検索の手法と頑健な自然言語処理を融合した高速、大容量の文書処理技術が急速に実用化されつつある。中でも明らかなのは、インターネット利用者にとっての高性能な検索機能の必要性である。検索機能の強化のために、文書内容間の類似性を高速に判定するソフトウェア・ライブラリを OS に標準搭載する動きさえみられる[Apple98]。本稿では、このような類似文書検索の一つである ConceptBase の技術概要を紹介する。ConceptBase では、統計手法による伝統的な情報検索の手法の１つであるヴェクトル空間法を複合名詞句間の処理によって洗練させ、データ疎問題の解決や、高速化をはかっている[Evans96]。さらに、業務上さまざまな局面で利用する文書群のスケールを上下に拡大するための新しいソリューション（業務上の問題解決）として、自動分類機能と、複数文書の鳥瞰ビューを提供する自動要約機能を紹介する。

高性能な検索機能が切望されている状況について、[Bannan97]は次のように描く:

"One large consulting company admitted that they could locate only 20% of their electronic files for their company products. We're not even talking about memos here."

何百万頁もの文書を生産することを生業としている会社で、必要な文書を適時に発掘、再利用できている率が２０%、という事例である。いくら電子文書には重さが無いし(あまり)かさばらないとはいえ、手作業による分類整理、検索の破綻した現場は多数派を占めている可能性がある。

この問題の解決案としては、WWW(World Wide Web)の技術標準に基づき、セキュリティを確保したイントラネット上で文書を統一的に管理するアプローチが有望、との見方が定着している。このようなイントラネット文書管理サーバーの実用性を左右する鍵としては、高速性、大容量性、そして、高精度の文書検索エンジンの有無があげられる。

検索機能が充実してきたならば、次の段階として、情報の構造化、そしてそれを踏まえた情報提示技術への必要性が高まると予想されている[武田 98]。インターネットの大規模検索エンジンは、高速、大容量の観点では十分過ぎる程の性能をもつに至っている。例えば AltaVista という検索エンジン[Selzer97]では、数個のキーワードの AND 条件検索を実行しただけで、数１０万件が該当した、という回答がただちに返ってくることがある。このような大量の検索結果の全貌をユーザが数秒以内に把握し、思考の流れを途切れさせずに、自分にとっての重要度を判定できるとは期待できない。この意味で、現在のインターネットの大規模検索エンジンの高速、大容量性は、人間の認知能力、GUI 操作能力の限界を超えてしまっている。その結果、大量文書の分布状況の視覚化などの情報提示技術や、複数の関連文書群から情報構造を抽出したり、複数文書を一括して高精度に要約してくれる機能が強く求められている。

優れた要約機能はインターネット利用の際に必要不可欠という認識が、一般にも浸透しつつある。Web 文書サイトの提供者としての心構えを説いた[ホームページクリエーターズフォーラム 97]では、Web 文書の読者が量の多さを喜ばないことを説明している：

テキスト中心のデザイン

■一方向に読み進んでいく読み物風なのか？

Web が一般の書籍と大きく異なるのは、ばらばらと斜め読みができないということです。ページをブラウズするごとに、ページとページの間には大きな「時間の壁」があるのです。　・・・（中略）・・・

書籍は買ったら読もうと努力しますし読まないと損をしたような気になるものです。ところが、Web は通話料金やプロバイダへの課金が価値の基準となりますので、ちょっと退屈するとすぐ他のサイトに行ってしまいます。この特性の違いは非常に大きいと感じます。

以上述べた背景のもとに、本稿では、以下、第２節で ConceptBase の言語処理の概要、第３節では、ConceptBase Search が特徴とする類似検索が何故有効であるか、従来型の Boolean 情報検索パラダイムと比較して論ずる。関連性フィードバックの形態の違いと、それに起因する「文書一覧の可読性」の重要性について指摘する。第４節では ConceptBase 関連技術の新展開として、１度に扱う文書空間のサイズを上下に拡大し、データベース、選択文書を構造化する、自動分類およびサマライザ技術について述べる。

2 ConceptBase の言語処理

ConceptBase が自然言語処理を情報検索に応用せんと試みた際の主要な目標、技術課題は次の３つである[Evans96]：

-1 大量の生テキストを高速に処理できること

→故に、１秒に１，２文しか解析できないような構文解析モジュールなどは採用しない。

-2 特定の分野、スタイルに依存しない非制限テキストを対象とすること

→実在する生テキストを幅広く扱えること。でき

れば OCR 認識誤りを含むテキスト等をも頑健に処理し十分実用的な解析精度、検索精度を達成すること。

-3 '浅い' 言語理解を実現すること

→統計手法の大きな弱点、「データ疎問題」は、1文書の内部でも、高々数万文書位の文書データベースにも、様々な形で存在する。その一解決方法は、普通名詞間の照応のような、同一対象のことを記述していながら異なる文字列で表現された名詞句間の関連付け処理を行うことである。このような '浅い' 言語理解により、実在の文書中で様々に言い換えられる同一の指示対象(Concept)を結びつけ、ヒストグラムを充実させることができる。

1.の要請を少し具体化すると、ギガバイト級の生テキストから転値ファイル・インデックスを高々数時間のオーダーで抽出し、高い精度を実現すること、となる。このため個々のインデックス内容を抽出する処理手順は図1([Evans96]を改訂)のような構成をとることになった。

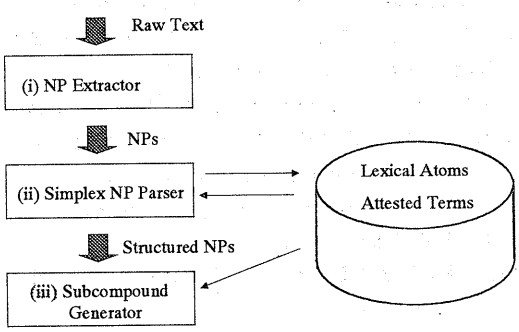

図1 ConceptBaseのインデックス作成前処理
～Evans96より

いずれも線形時間で計算する、(i)高速な形態素解析による名詞句区間の同定、(ii)名詞句の分解処理、(iii)複合名詞句間の構成語の部分マッチング(対象文書中の名詞句を分解した蓄積データを参照しつつ登録)という逐次処理からなる。

分野依存、特定の文体などの知識ベースにまったく依存していないことで、2.の要請を満たしている。また、かつての AI 研究の流れで情報検索に自然言語処理を応用した PLANES, FRED, LIFER などのｼｽﾃﾑ構成[Findler91]と比べるとはるかに単純である。にもかかわらず、(iii)によって 3. '浅い' 言語理解の実現、という目標もある程度、達成することができている。

3 ConceptBase Search のソリューション

1997 年にﾘﾘｰｽされた ConceptBase Search という類似文書検索ソフトウェアは、図2に示すような検索結果出力画面等の文書検索 GUI(Graphical user Interface)をもつ。ConceptBase Search では、予め文書データベースごとに作成した転値ファイルをもとに、指定した文書データベース(複数)の全文書とクエリー(基準文書)との類似度を判定する。その結果、上位 N 文書(N の既定値は 50)を表示し閲覧に供する。これが基本的な動作である。

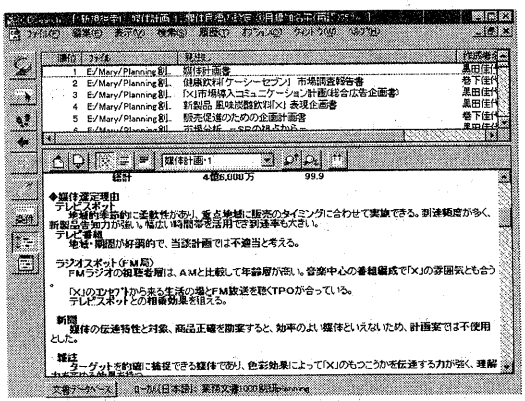

図2(a) ConceptBase Search の検索結果画面例

図2(b) ConceptBase Search Web Client の画面例

ConceptBase Search の検索エンジンは単体の Windows マシンに置くこともできるし、TCP/IP の LAN で結ばれたサーバ上に置いて利用することもできる。Windows アプリケーションの専用クライアント(図2(a))と、汎用の Web ブラウザ上で利用可能な Web Client(図2(b))の2種を提供している。両者合わせて、製造業、流通・サービス業、公共・金融業を中心に3万本程度が幅広く使われている。Windows 版の専用クライアントと、Web Client の利用比率は2対8程度である。

3－1 類似文書検索の必然性と ConceptBase Search

図2に示した ConceptBase Search 以外にも、類似の検索 GUI をもった類似文書検索ソフトウェア製品やサービスが登場している。[野村 97]の巻末にまとめられている以外で最近、幅広く使われ始めた製品がある。MacOS8.5 にバンドルされた Sherlock である[Apple98]。Sherlock では、週次のスケジューリングに従ってファイルシステムのボリュームごとに転値インデックスを作成しておき、OS 標準のファイル検索の拡張機能として高速な類似文書検索を実行可能にしている。さらに従来通りの文書名を中心とした文書書誌情報の検索やインターネット上の検索エンジンの活用機能が同じダイアログの隣の'タブ'で呼び出せるようになっている。

このように、「類似文書を同時に一括してランキング」するスタイルが、インターネットを前提とした電子文書環境、情報空間の中で1つの「検索」パラダイムとして認知されつつある。これは何故であろうか。1つの理由は、文書データベース構築の際に予めキーワード集合を人手で付与する必要が無いことであろう。キーワード集合のアプローチの限界として次のようなポイントが指摘される[野村 97]：

- そもそも文書を予め少ないキーワード集合で抽象化し特徴付けておくことはできない；
- 「できないことは苦痛である」→キーワード設定は苦痛；恣意性大；故に実行されない；
- キーワード集合は、視点や興味の焦点の違いにより無効となりがち； 同一の執筆者、読者であっても時間や作業文脈とともに視点は移り変わる（「1週間後の自分は赤の他人」との自戒）

別な角度から理由を考えてみると、たとえば以下のような論点がみつかる：

- 情報空間をある視点で切って眺める目的で文書集合を構造化するには、文書間の近さ、遠さを表す何らかの評価尺度、即ち類似判定が必要；
- 文書集合から部分集合を絞り込む(='検索','分類') ための手がかりとしても上記尺度が必要；
- 文書間に共通する、もしくはその差違を特徴付ける一部の属性を'抽出'するのにも類似判定が必要；
- 情報空間へ人間が参加した際、過去の経験(=検索履歴)との類似判定によって共通点や差違をシステムが際だたせてくれると、人間の高度な連想能力が有効に発揮されるようになる；

類似性や差違の活用が有効なのは認めたとしよう。類似性を判定するのに、文書の本文でなく文書登録の際に 5W1H の属性を記述した書誌情報や、既存の RDB を使えば確実ではないか、という見解があるかもしれない。これに対しては、書誌情報も「キーワード集合」と同様、予め最適に設定出来ているとは限らないし、「キーワード集合」ほどにも本文の内容を反映していない、と反論できる。加えて、企業情報システムなどの実務の現場では、管理者や自動登録ｴｰｼﾞｪﾝﾄがまとめて同じ日時に同じ文書見出し情報(例えば『営業日報』)で、登録した結果、書誌情報が無効となることもある。文書の登録内容が更新された際に 5W1H の情報が上書きされたりする運用上の問題点にも注意が必要である。

以上により、文書の本文内容に基づいて全自動で類似性を判定する技術が、大量に電子文書の氾濫する環境における問題解決の主軸となるべき必然性が浮かび上がってくる。

一旦、文書の本文内容主導の(= contents driven)情報システムを構想し、設計することに決めたならば、ユーザーにとっては、キーワード集合や書誌情報などは極力見えないようにするのが望ましい。本文内容に思考を集中し、いわゆるホワイトカラー業務の知的生産性を向上させることがビジネス現場でのソリューションに直結するからである。

この設計思想に沿って、ConceptBase Search では、以下の観点で接続性の拡大をはかっている：

(1)多種多様な文書フォーマットをユーザーが意識せずに扱えること。

(2)LotusNotes DB を含むバックエンドデータベースを仮想化し検索結果の出自を意識せずに利用できるようなゲートウェイ機能を充実させること。

3－2 ConceptBaseSearch の特徴のまとめ

前節では、類似文書検索の必然性について考察し、ConceptBase Search の接続性拡大の設計思想を導いた。ここではまず、キーワード指定による Boolean 検索や、いわゆる全文検索システム等と対比しながら、情報検索分野でユーザが伝統的に抱えていた問題点へのソリューションに直結する形で、ConceptBase Search の特徴をまとめる。

特徴1　クエリーの長さに制限の無い自然文入力
→誰でも使える（cf. Boolean）
長文や単語の羅列で指定可能（and,or での指定不要）

特徴2　類似度付きで判定結果を一括表示　→素速く目的の情報に到達できる（cf. 全文検索）

・全文書を細かくランキングすることで関連の強い順に結果表示
・存在が既知でない文書の発見
・絞り込み型の検索で起こる検索漏れの防止

特徴３　関連語自動検出　　→柔軟で高精度な検索
・背景知識や先入観に影響されない最適な関連語を検出し自動獲得シソーラスとして利用

特徴４　全自動インデキシング →導入・運用コストを削減できる
・DB 設計不要（DB 作成の自動化）
・文書分類不要（文書登録の自動化）
・再現率を高めるためのシソーラス辞書不要
（異表記対応等で併用は可：正解率低下を回避できる範囲で推奨）

ConceptBase Search では、上記「特徴２」をさらに有効なものとするために、視覚化にも一定の配慮を施している。関連度の横棒グラフや文書プレビュー画面の提供、さらにプレビュー画面中で、クエリー中のターム（複合名詞句）との類似タームを青色で強調表示したり、高頻度の類似タームを多く含む文を強調表示したりすることにより、文書内容を素早く把握できるようにｻﾎﾟｰﾄしている。さらに、原文書へのパス名を文書データベース中に保存すること、および各種文書表示プラグ・インの提供により、必要に応じてを原文書のレイアウトやカラー・デザインそのままの表示をさせることができる。

３－３ 関連性フィードバックの新しい位置付け

図２に示した検索結果の文書一覧画面において、文書内容のプレビューを閲覧したりしながら、欲していた文書をいくつか反転指定する。その上で再度検索を実行すると、これらの文書から抽出した全てのタームの重み付きヴェクトルとの類似度判定によってデータベース中の文書が改めてランキングされ、文書一覧が更新される（図３）。このような機能は伝統的に関連性拡張、適合性フィードバック、あるいは関連性フィードバック(Relevance Feedback)と呼ばれている。

なぜ関連性フィードバックが情報検索のソリューションとして重要であるかについて、[NTT-IT97]を引用した[野村 97]から再録し確認しておきたい：

通常、ユーザは検索したい概念を単語や文として最初から的確に表現することができません。InfoBee/TR では、適合フィードバックという方式を用いたユーザとの対話的なインタフェースによって検索漏れの少ない情報検索を行うことができます。

まず、思い付く単語によって検索すると、結果は適合度順に表示されます。次に、１位から１０位くらいの結果の文書を調べてみて、自分の要求するものに近い文書の タイトルの左側の四角いボタンをチェックし、類似文書検索ボタンをクリックします。ＩｎｆｏＢｅｅ／ＴＲはこれらの文書中に含まれる重要度の高い単語を使って、再度、検索を行い適合度順に結果を表示します。すなわち、文書に含まれる重要な単語の集合によって概念を表して検索を行うことができるのです。

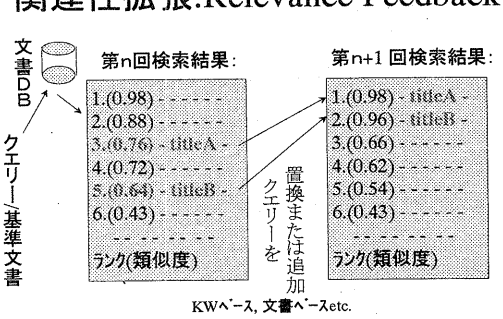

図3 関連性フィードバックのユーザインタフェース

仮に、上記の「通常」の場合に反する稀なケースとして、ユーザが検索の結果として得られるべき文書の集合の明確なイメージをもっていたとする。この場合でも、ユーザが予め知り尽くすことのできないデータベースの構造や内容の充実状況によっては、検索結果を修正、改良して欲しくなることがあるだろう。この際に最も簡便なユーザインタフェースとして考えられるのが、より適合度の高かった検索結果中の文書をクエリーにして検索を再試行すること。すなわち関連性フィードバックである。

プレビュー画面や、あるいは、マルチメディア・オブジェクトの貼り込まれた原文書のレイアウトそのものを提示することによって、より適切にクエリーとすべき文書を選べるような期待もある。この点を含め、伝統的な情報検索における関連性フィードバックと類似検索における関連性フィードバックとを対比した分析を確認しておきたい。

図４(a)は、[Velez97]で図解された伝統的な関連性フィードバックのパラダイムである。検索結果ヒットした文書数を主たる判断基準にして、それが多過

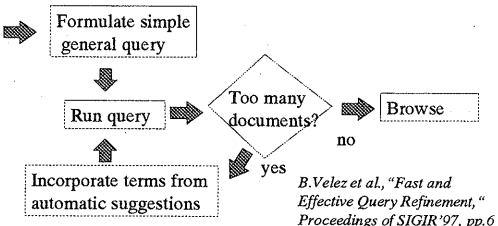

図４(a) 関連性フィードバック -Boolean

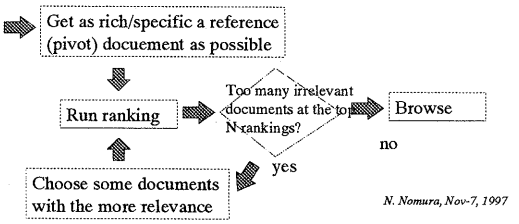

図４(b) 関連性フィードバック -類似検索

ぎたら検索式に AND 条件を付加し、逆に少なすぎたら OR 条件を付加するなどして検索式を修正する。

これに対し、図４(b)に図解した、類似検索における関連性フィードバック[野村 97]では、Boolean 検索式をこしらえる代わりに、要求内容になるべく類似した文書を用意する。内容を詳述していればいるほどよい。１文書だけでなく、内容が関連してさえいれば、できるだけ多数の文書を用意しても良い。

データベース中の全文書との間で類似判定を実行すると(Run Ranking)、必ずランキングの結果が返ってくる。このため、初期クエリーの段階からクエリー長ければ長いほど良い、というのが Boolean 検索との大きな違いである。関連性フィードバックの際にも、図３(a)の状態に見えたら図３(b)の状態に近づけるように、即ち無関係の文書が文書一覧から減るように、好ましい文書を指定する（ちなみに ConceptBase では無関係過ぎる文書を指定してマイナスの重みを与えることも可能である)。

図４の対比から、類似文書検索におけるユーザインタフェース上の要求項目がいくつか導き出される。第一に、図４(b)中央にある判断条件「無関係な文書が上位に多く現れ過ぎていないか?」が有効に機能するよう、文書一覧中の文書見出しの可読性、了解性が高くなければならない。次に、これらの文書見出しに準ずるプレビューや要約結果が、短時間で原文の大意を忠実に汲み取れるような質の高いものである必要がある。この他、ユーザが、「この辺りが潮時。これで十分。」と感じやすいように、関連性の高い文書が一定量以上、稠密に集まったことを感得させる視覚化（色、棒グラフなどの表現）なども、類似文書検索の使用感に大きな影響を与える可能性がある。

4　文書空間の構造化をもたらす新技術

本節では、類似文書検索を大規模に行って、各部署へ検索・分類の結果を配信するという応用を念頭においた自動分類機能の一例 CB Classifier と、検索結果の内容の一覧性を強化するサマライザ機能（複数文書内容の構造化と一括要約）について紹介する。

4－1　自動分類機能 CB Classifier

ヴェクトル空間法[Salton68]の自然な拡張として、多数のターム・ヴェクトルの和集合演算と加重平均をとったヴェクトルをその文書集合を代表するプロファイルとして活用することが提案されている。この文書集合の帰属先をデータベースとみなせばこれはデータベースのプロファイルとなる。同様に、個人プロファイル、部門プロファイル等を作ることができる。

文書の分類カテゴリーにプロファイルを与えれば、自動分類機能としてまとめあげることができる。多数の文書から抽出したターム・ヴェクトルを処理して作成したプロファイルに分類カテゴリーの特徴を代表させることによって、高々分類数と、被分類対象の文書数の対数との積に比例する程度の小さな計算量で大規模な自動分類を実現できる。この機能のプロトタイピング結果の一例を図５に示す。

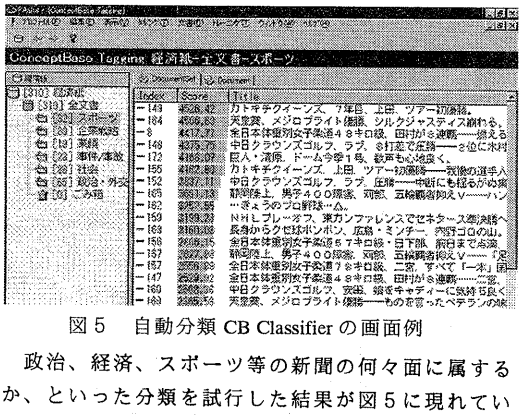

図５　自動分類 CB Classifier の画面例

政治、経済、スポーツ等の新聞の何々面に属するか、といった分類を試行した結果が図５に現れている。類似検索の場合と同様に、ここでもユーザが自動分類の結果を訂正したり他の条件による自動分類結果と付き合わせることにより関連性フィードバックを行うことができる。

4－2　CBサマライザ

従来製品化されていた要約機能には、ad-hoc(場当たり的)な規則の集積によるものや、統計手法を応用して１文書内で一種の検索を実行し文の重要度を数値化して文選択を行うものがあった（他の様々な手法については[奥村 98]を参照されたい)。後者のアプローチ、例えばヴェクトル空間法による検索の副産物としての要約機能[野村 97]では、章題目のように重要な短いフレーズを落として読みにくい長文ばかり残す、といった問題点がある。「高頻度に現れる名詞句を多く含む文を残す」という原理で制御しているのだから当然である。また、構文解析や文脈解析を行わない結果として、代名詞類や、日本語の場合「こそあど」を全く考慮せずに文やフレーズを省略することとなり、可読性、了解性、そして原文の文意への忠実性に乏しい結果が頻繁に生成されていた。

ＣＢサマライザでは、これらの問題点を解決するために、文書レイアウトの解析結果や、「こそあど」を初めとする照応解析から、無為に切り離してはいけない文、フレーズの組み合わせを制約として抽出する。たとえば、文書レイアウト解析によって推定した主見出しに副見出しを従属させたり、一続きの箇条書きを束ねる。これらの制約は、指定された要約率に基づいてフレーズを取捨選択する際に適用する。箇条書きならその一部をだるま落とし(例：第１条、第３条を残して第２条を落とすなど)することなく、必要ならば各条を短くする、という制御を行う。「こそあど」で関連付けた先のフレーズ、文、段落についても、もし先行詞相当部分が長大で落とさざるを得なくなった場合には、関連付けのきっかけとなった「こそあど」を含むフレーズを適宜切り落として誤読を回避するように制御している。低レベルの関連付け処理としては、開き括弧と閉じ括弧の対応や、引用文、補文の範囲の推定に基づく制約抽出も実行している。これらをまとめると、談話の結束性(cohesion)を破らないような制約を抽出し、要約結果をつないで*読*めるよう制御したシステム、といえる。

また、論旨の流れを、汎用的な談話文法のルールセットによって捉えようとしている。ここでは、接続詞、接続助詞、陳述の副詞、提題助詞・助動詞類などを手がかりに上述の関連付けに準じた形でフレーズ間の依存関係を判定する。逆接の接続詞と提題、陳述の手がかりを結びつけることにより、文書全体の主題へ接近する陳述か離反する陳述かを弁別し、相対的な重要度を判定したりする処理を行っている。

要約内容を具体的に制御するパラメータとしては、(1)段落、章・節、文書全体の先頭や末尾に近いところに主題文が出現することを推定する度合いを制御するパラメータ、(2)陳述を含む部分を重視するか高頻度タームを重視するかを切り換える「意見 vs 事実優先」パラメータ等を提供している。

要約結果を生成するフェーズの上流工程では、重文、複文から抽出した文の骨格（提題と陳述を中心としたもの）を抜き出す処理を行う。この処理は、重文、複文の構造解析に特化した構文解析のルールセットを用い、独自の決定性構文解析エンジンにより毎秒数百文程度の速度で実行している(Pentium200MHz)。解析フェーズにおいて総合的に与えた重要度と、各種関連付けの制約、それから当該の文自体の構文的複雑さを判断条件にして、要約結果に文全体を残すか、従属節を落とすか、さらには骨格以外を全て落とすかを決定している。

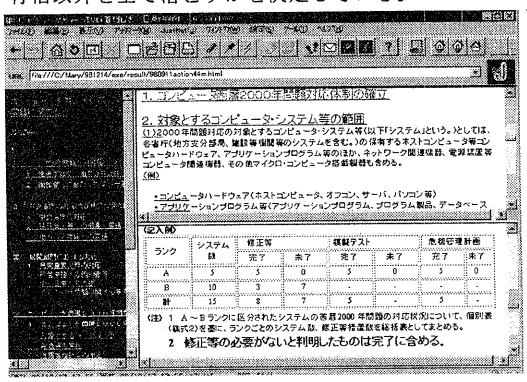

図６　CB Summarizer の出力画面例

最終の生成結果は、プレーンテキスト出力または、HTML 出力、あるいはその両方を同時に出力することができる。HTML 出力では、原文書のレイアウト構造、論理構造を解析した結果に基づいて作成した「目次フレーム」、重要箇所を一部大きめのフォントなどで強調表示した「要約フレーム」、そして、複数の原文書を結合し、そのレイアウトの一部を再現して整形した「全文フレーム」とを生成する。図６の左側、縦長の「目次フレーム」の全ての行から「要約フレーム」中の当該部分へハイパーリンクを張っている。「要約フレーム」から「全文フレーム」中の当該部分へもハイパーリンクを張ると同時に、「要約フレーム」中に採択された文やフレーズは「全文フレーム」中で強調表示している。この強調表示のデザイン（フォント種別、サイズ、色などの文字飾り）は、設定オプション中の HTML タグのサンプル（デフォ

ルトは太字を意味する<B>）を書き換えることで容易に変更可能である。

このようなフレーム構造をもった出力結果の最大の利点は、文書（群）全体を鳥瞰しつつ、興味をひいた箇所について即座に詳細情報を段階を追って読めることである。多くの HTML ブラウザが、リンク元を一度クリックしたら色を変えて表示してくれることから、どのポイントに先刻目を通したかが視覚的に瞬時に把握可能となる。活用の仕方しだいでは、紙のページをぱらぱらとめくるよりも高性能な「読む」環境を提供することができる。実際、「読んでいて印刷物や電子書籍よりも安心感がある。長いテキストデータは全てこのサマライザを通して読みたい。」という体感評価の声が寄せられている。

一方、[奥村 98]にまとめられている要約技術開発の流れの中では、要約結果を indicative なもの、すなわち原文書を読むか否かを決める手がかりに過ぎないとするか、informative なもの、すなわち原文書を捨てて要約結果が一人歩きすることを前提とするかの議論が指摘されている。この観点からＣＢサマライザを眺めるなら、両者を融合し、読者の必要に応じてミクロに選びながら１つの要約結果を活用する形態を提供した、と評価することができる。

なお、HTML 出力中には、フレーズごとの重要度やフレーズ間の関連性等の解析結果をフォントデザインやグラフィクスの併用、特別な相互リンク、等によって反映することができる。従来の文章要約のパラダイムがもっていた「重要度を０か１かに正規化して落とすか残すかのいずれかに決定」という制約から要約システムを解放し、さらに文章の論旨の流れを視覚的に浮かび上がらせて表示するような可能性を秘めている。現在は、図６に示した構造を自然に利用できるように、解析結果得られた多くの情報の表出を抑止しているが、将来は、文章推敲のための支援ツールに特化させるといった方向性も検討していきたい。

5 おわりに

以上、ConceptBase の言語処理と新しいソリューションについて紹介した。ConceptBase のコアの拡張に関する今後の見通しとしては、性能面では、ギガバイト級からテラバイト級へのスケールアップや文書インデックス作成(Database build)の高速化が課題として上げられる。精度面では、正解率の向上を狙った深い言語解析結果の活用や、再現率の向上を狙った既存シソーラスの「ほど良い」適用、クラスタリング技術の併用などを検討していく必要がある。

一方、アプリケーションとしては、ユーザ管理、スケジューリング、障害対策を含むサーバー機能の充実をめざしている。さらに、文書配信機能との連動、ワークフロー連携などを可能にすべく既存の文書管理サーバーや RDB、ODB との接続性の向上をはかっている。クライアント側においても、Justsystem Office9 以外のクライアントアプリケーションとの親和性の拡大、などを考慮していく必要がある。

今後は、ConceptBase 技術が電子文書アクセス環境の初心者や、電子掲示板アクセスの新参者に優しい、より「民主的な」情報アクセスを支援するインフラ構築への貢献できるよう開発していきたい。そのためには、実用的な言語処理技術と各種情報検索技術との融合への傾斜を強め、新技術を利用の現場で着実に評価し、技術開発にフィードバックする必要がある。それとともに各種標準との接続性をさらに拡大し、常に現場で評価される環境を整えていきたい。

参考文献：

Apple98: MacOS8.5 , Apple Computer Inc., 1998

Bannan97: "Intranet Document Management -- A Guide for Webmasters and Content Providers," Chapter 9.. Managging Large Collections of Documents, pp.169 "Case History 1," Addison-Wesley, March, 1997

Evans96: Evans,D., and Zhai,C. "Noun-Phrase Analysis in Unrestricted Text for Information Retrieval, " in Proceedings of 34th Annual Meeting of the Association for Computational Linguistics, pp.17-24, 1996

Findler91: Findler, N., V. ed., "An Artificial Intelligence Technique for Information and Fact Retrieval," MIT Press, 1991

NTT-IT97:InfoBee/TR(NTT ｲﾝﾃﾘｼﾞｪﾝﾄﾃｸﾉﾛｼﾞｰ株式会社) http://www.iijnet.or.jp/ntt-it/goods/1ji/infobee.html （1997 年末当時）

Salton68: Salton, G., "Automatic Information Organization and Retrieval," New York, McGraw-Hill, 1968

Selzer97: Selzer, R., Ray, E., Ray, E. "The AltaVista Search Revolution," McGraw-Hill Companies, 1997 （邦題「AltaVista 完全活用ガイド ～インターネットのすべてを検索する方法」,翔栄社刊）

奥村 98:奥村学,難波英嗣「テキスト自動要約に関する研究動向」, 電子情報通信学会 NLC 研究会講習会「自然言語処理と情報提示技術」資料 pp.1-24,1997.11.12

野村 97: 野村直之「言語工学による類似情報抽出の精度向上とその応用」, 電子情報通信学会 NLC 研究会講習会「自然言語処理と検索技術」資料 pp.33-56,1997.11.7

ホームページクリエイターズフォーラム 97: ホームページクリエイターズフォーラム編,「ＨＴＭＬの次はこれだ！ ～ホームページ改造大作戦」富士通ブックス, ISBN4-938711-89-3