システム開発支援のための情報共有システム
Information Sharing System for Supporting System Development

国立情報学研究所 実証研究センター
高須　淳宏

システム開発において、そこに携わる技術者の間で情報を共有することが、信頼性の高いシステムを作り運用する上で重要な要件となる。しかし、システムの複雑化は、そのシステムに関する情報量の増加を招き、また、システムの開発・運用に携わる技術者数の増大を招く。その結果、情報共有はますます難しくなり、新たな情報共有の枠組が必要になっている。本稿では、複雑なシステムを設計、管理、運用するために、多様な場所に多様な形態で蓄積された情報を統合的に活用するための情報共有システムMIMIRを紹介する。システム開発に必要になる情報を技術者間で共有するためには、Web等で公開されている情報を効率良く収集利用するとともに、技術者の頭の中に留まりがちな情報を引き出し、電子化することが重要になる。しかし、文書化作業は技術者にとって優先度の低い作業である。そこで、本稿では音声データやセンサデータなどを直接用いることによって、技術者の文書化コストを下げる技術を中心に筆者等の研究を紹介する。

1．はじめに

現在の社会では、自動車や航空機などの物理的なシステムからコンピュータソフトウェアによって実現されるサイバーシステムまで多種多様なシステムが人間の社会活動を支えている。これらのシステムは、新しい技術の導入や改良を繰り返すことによって、より便利な機能を備えていく一方で、その複雑さは増大の一途を辿っている。

システムの開発においては、そこに携わる技術者の間で情報を共有することが、信頼性の高いシステムを作り運用する上で重要な要件となる。しかし、システムの複雑化は、そのシステムに関する情報量の増加を招き、また、システムの開発・運用に携わる技術者数の増大を招く。その結果、情報共有はますます難しくなり、新たな情報共有の枠組が必要になっている。

システムに関する情報は仕様書や設計図のような形で文書化され共有されるものと、技術者個人や一部のグループに留まるものがある。そのため、技術者がシステムに関する情報を入手する場合に、仕様書等の文書から情報を入手するとともに、その情報に詳しい技術者から直接話を聞いて情報を入手することも必要になる。前者は、システマティックな情報共有の枠組となっているのに対して、後者は、複数の人間が分担して情報を記憶し、人間のネットワークを介して情報を取得する情報共有の枠組となっている。このような人間を核とした情報共有の枠組は、人間の高度な情報処理能力ゆえに、柔軟な情報の取扱いが可能になる一方で、膨大な情報を共有することには適さず、また、人間がいなくなると

情報も失われてしまうため、永続性、信頼性の観点からは必ずしも優れた枠組とはいえない。上記２種類の枠組によって共有される情報の比率は分野によって異なるものと思われるが、例えば、知識労働者が扱う情報の80%が共有されることなく個人に留まり、個人の移動とともに失われてしまうという調査報告もなされている。

システムの複雑化は、共有すべき情報量の増大、個々の情報の相互関係の複雑化、共有すべき人間の増加につながり、情報共有を難しくする。特に、人間を核とする情報共有の枠組においては、システムに係わる技術者のネットワークが複雑になり、誰に聞けば必要な情報を得ることができるかがわからなくなり、情報を取得することが非常に難しくなる。また、近年は、組織における人間の流動化が進んでおり、必要な情報を保有する人間がわかったとしても、組織外への移動に伴い、その人にアクセスすることが自体が難しいという状況もおこる。そのため、今後は情報を電子化し、コンピュータも活用した情報共有の基盤を構築することが重要になる。

情報を電子化して共有するためには、情報共有システムへ情報を入力するプロセスと情報共有システムから効果的に情報を取り出すため検索プロセスがシームレスにつながることが重要になる。システム開発においては、システム自体の開発に技術者の力点が置かれ、情報共有のための情報の整理と電子化は後回しになりがちである。そのため、情報共有システムには、技術者が情報を電子化するためのコストを低減する機能が求められる。また、システム開発で扱われる情報は、その内容、記述形式、記述メディアなどさまざまな点で多様かつ不均質であるため、その情報を活用するためには、高度な検索技術が必要になる。さらに、現在は、インターネットの普及によって、各種の情報がネットワーク上に蓄積され、巨大な情報源が形成されている。そこには、システム開発に有用な情報も多く含まれているため、システム開発チーム内部の情報共有とともに、これらの外部情報とのリンケージも情報共有システムの重要な機能になる。本稿では、これまでに筆者等のグループが行ってきた研究を中心にシステム開発・運用において技術者が情報を共有するための情報共有システムMIMIRを紹介し、そこで取り組んできた要素技術について述べる。本稿では、特に技術者による情報の電子化コストをさげるため、音声やセンサなどの生データを活用する方法に焦点を当てる。

２．情報共有システムの概要

本章では、筆者等が開発を進めている情報共有システムMIMIRの概要を示す。MIMIRは人工衛星等のシステム開発および運用過程で技術者が収集・生成する各種情報を管理することを目的としている。特に

- システム開発運用の各フェーズでの技術者による情報共有支援
- システム開発運用のフェーズの間での情報の継承支援
- 類似システム開発のための情報のアーカイブ構築

を主たる目的としている。図１はMIMIRの概要を示す。

システム開発の段階で技術者はさまざまな情報を扱うことになる。MIMIRが扱う情報は、個人情報、開発運用チームで共有する内部情報、および、一般に公開されている外部情報の大きく別けて３種類に分類される。個人情報は電子メール、メモ、ブックマーク等、技術者個人が管理する情報である。内部情報は、技術者等が開催するミーティングの議事録、配布資料や開発段階で行われる各種実験の報告書等、情報共有するために技術者等が作成する各種文書より構成される。しかし、

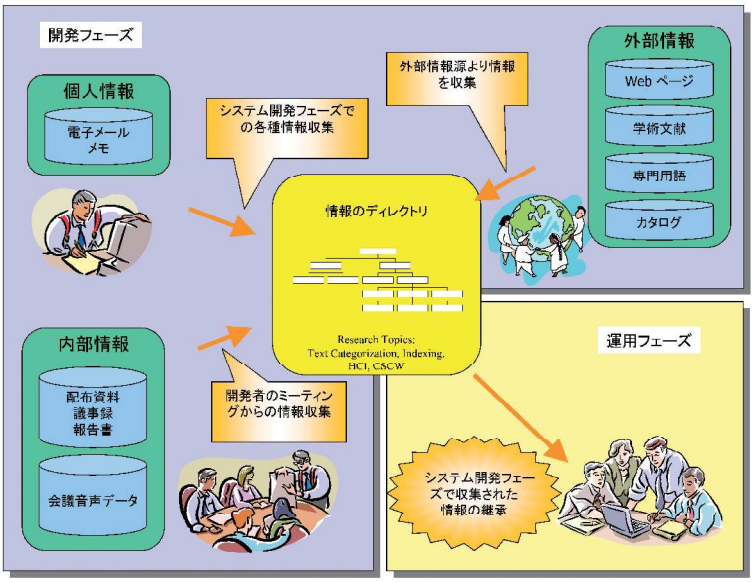

図１　情報共有システムMIMIRの概要

技術者によって文書化される情報は非常に限られているため、MIMIRは会議における技術者の音声データや実験における各種センサデータも内部情報として蓄積する。外部情報としては、システムの設計に必要な部品のカタログや学術文献などへのリンクを扱っている。近年はWeb上に非常に多くの情報が蓄積されており、Webから必要な情報を得られることも多い。そこでMIMIRはWebを貴重な外部情報源として扱い、Webページに対するリンク情報も外部情報として扱う。

蓄積された情報を利用する観点からは、これらの情報が独立に格納されているだけでは、必要な情報を取り出すことが困難になる。MIMIRは、下記の２種類の情報アクセスを提供する。

- ディレクトリを用いた俯瞰的情報アクセス
- 情報検索による網羅的情報アクセス

システムの設計開発においては、システムは通常モジュールに分割され、モジュール単位で技術者が活動することが多い。また、複雑なシステムでは、モジュールはさらにサブモジュールに分割され、階層的な構造を持つことが多い。この階層構造は、情報共有の対象となるシステムに対する技術者の共通の構造となる。そこで、MIMIRでは、システムの階層構造をベースにして情報を管理する。MIMIRで扱われる情報は、階層構造を持った情報のディレクトリの各ノードにリンクされる形で格納される。ディレクトリ構造は、MIMIRが持つ情報を俯瞰する上で重要なデータ提示方法となっている。一方、MIMIRが保有する情報の中から必要な情報を網羅的に探す手段も必要になる。このような場合は、すべての文書に対して、検索語を用いた情報検索を行う。MIMIRに格納される情報は基本的にテキスト情報であるが、音声やセンサーなどの非テキスト情報も含まれている。これらの情報を検索するために、個々のメディアに適した情報検索に関する機能についての研究も進めている。

このような情報共有システムを実現するた

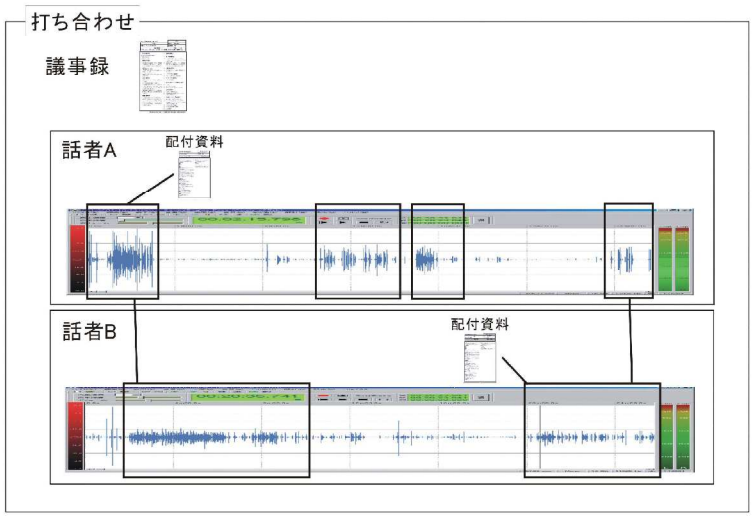

図２　議事録と音声データの対応づけ

めには、さまざまな情報処理技術が必要になる。外部情報と共有システム内の情報をリンクするためには、Webに対するfocused crawlingや情報のマッチング技術［９］が必要になる。また、内部情報に関しては、音声データやセンサデータの利用のためには、信号処理技術に加えて多メディアに対する検索技術が必要になる。さらに、設計に関する情報をディレクトリにまとめるためには、文書クラスタリングや文書分類［４］の技術が中心的な役割を果たす。情報共有システムの開発で、筆者等はこれまで、文書分類技術や近似マッチング、センサデータに対するアノテーション付与に関する研究を進めてきた。本稿では、これらの研究のなかで、非テキストデータを利用するための技術を中心にこれまでの研究を紹介する。

３．近似文字列検索

会議などで得られる技術者の音声データを活用するためには、音声認識技術を用いて音声データをテキスト化し、テキストに対する情報検索技術を用いて情報を取り出す方法が考えられる。近年、音声認識技術は着実に進歩しており、高い認識精度が得られるようなってきている。また、音声データに対する情報検索法の研究も進められている[8,5]。しかし、録音時のノイズや音声認識技術の精度の問題から、現段階で有効な情報検索を行うことは非常に難しい。

そこで、MIMIRでは、音声データに対して２種類のアクセスパスを設けた。ひとつは、会議の議事録と議事録の各議題に対応する音声データの対応づけを行うことによって、議事録の詳細データとして音声データを使うというものである。システムの設計･開発では、技術者は会議を行い問題点を共有したり、解決法についての議論を行う。このような会議の内容は、議事録のような形で文書化されるが、議事録に記述される内容は限られている。そこで、会議の音声データを録音し、文書化されたデータと音声データの関連部分の連携をとることによって、文書を経由して音声データを有効に活用することが可能になる。文書と音声データの関連付けは、一般の音声認識よりも容易な問題である。図２は、このシナリオを表しており、まず、録音した音声データから、話者単位で音声データを分

割し、さらに近似文字列マッチング技術を用いて、文書と音声データの関連部分を結びつける。

もうひとつのアクセスパスは、音声データ中に指定された単語が存在するかどうかを判定するword spottingを用いて情報検索を行う方法である。この方法では、事前に検索に用いられる単語のリストを作成し、それらの単語が音声データ中に表れるかどうかをword spotting技術を用いて調べるというものである。このようなことを行うためには、検索に使われる用語を予め用意しておく必要がある。技術者の会議などで用いられる用語には、その技術者グループに特有の表現があるために、技術者グループに適した専門用語集を構築することが望ましい。そこで、MIMIRでは、技術者が用語集を編集するための辞書編集システムを構築している。図3は、編集システムのインタフェースを示している。図に示されるように、用語集には用語とその説明に加えて、その用例、技術者グループ固有の表現と一般的な表現の併記、Web上の関連ページへのリンク、その用語に詳しい技術者へのリンクなどを含んでいる。これにより、word spottingに用いる用語集を構築するだけでなく、新人技術者がグループで良く用いられる用語を学ぶのを支援したり、その情報に詳しい技術者を表示することによって技術者ネットワークを構築することを促進することも意図されている。

音声データの検索では、音声認識の結果得られる文字列に対して近似文字列マッチングを適用することも考えられるし、音声認識の過程で得られる音素列に対して近似音素マッチングを適用する方法も考えられる。いずれの場合でも、シンボル列に対する近似的なマッチングが重要な役割を果たす。文字列の近似マッチングには従来から編集距離が用いられてきた。編集距離では、2つの文字列が与えられたときに、一方の文字列に挿入、置換、削除の編集操作を施すことによって、もう一方の文字列に変換することを考える。このときに、変換に必要な最少編集操作数を距離として用いる。編集距離では、編集対象となる文字には関係なく、編集操作のみで距離が定義されてしまう。しかし、音声認識のよ

図3　専門用語の編集

うな認識プロセスでは、文字によって誤り易さが異なるため、文字レベルで類似度を定義したほうが、より正確な類似度を計算することが可能になる。Confusion Matrix［2］は文字対に対して、その類似度を定義することによって認識器に即した類似度を与えることができる。

2つのシンボル列の類似度は、話者や会議が行われた環境によっても異なる。そのため、話者や録音環境に適応して類似度を定義することが望まれるが、編集距離の各編集操作のコストやConfusion Matrixのパラメタを求めることは、コストの高い作業である。そこで、これらのパラメタを訓練データから獲得する研究が行われている。Ristad等は編集距離のパラメタを訓練データから学習することが可能な可学習型編集距離を提案している［3］。しかし、このモデルでは、編集操作のレベルでしか類似度を定義できず、音声データに対する類似度モデルとしては不十分である。筆者等はConfusion Matrixのようにシンボルレベルで類似度を柔軟に定義でき、かつ、訓練データを用いて学習することによって類似度の定義を容易に変更できる統計モデルを提案した［6］。図4にこのモデルの例を示す。このモデルは、隠れマルコフモデル（HMM）の拡張となっており、モデルの各状態はある確率分布にしたがって文字列対を生成する。また、各状態で出力される文字列対の長さは固定されている。例えば図4の状態sは、各々長さ1の文字列対を出力する。各状態で出力される文字列対には確率が付与されており、例えば状態sにおいて、文字列対(a,a)は確率4／9で、また、文字列対(a,b)は確率1／9で出力される。HMMと同様に、ある状態から別の状態に遷移する確率も定義されており、例えば、図4では、状態sから状態rに遷移する確率は1／5となっている。このモデルでは、各状態で文字列対を出力しながら状態遷移を繰り返すことで、文

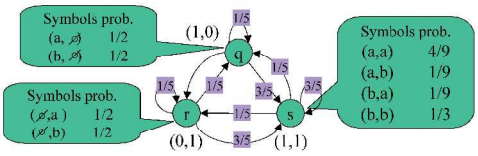

図4　文字列の類似度モデル

字列対（α，β）を出力する。また同時にその文字列を出力する確率も計算することができる。このモデルでは、この確率を文字列αとβの類似度として用いている。

図4において、状態qは長さ1の文字列と長さ0の文字列を出力することになり、編集距離における削除操作を対応している。同様に状態rは挿入操作に対応している。また、状態sは置換操作と無操作に対応している。一方、出力確率によって、文字レベルでの編集操作のコストを記述している。

4．センサデータの活用

近年、センサを用いた各種モニタリングシステムが普及したこともあり、センサデータ処理に関する研究が活発に行われている［1］。多くの物理的なシステムでは、状況を把握するための各種のセンサが取り付けられことが多い。例えば、人工衛星の運用時には、軌道上の人工衛星から送られてくる各種のセンサデータをもとに、人工衛星の状態を把握し、故障に対しては、限られた時間内にその原因をつきとめ、対策を講じるといったことが行われる。センサデータからシステムの状況を把握するためには、センサのパターンとシステムの状態を結びつけることが必要になる。

通常、システム開発では、さまざまな実験が行われる。これらの実験の中では、センサによって観測されたシステムの状態の分析が行われる。また、システム運用中にも、センサデータを分析してシステムの状態が推定される。実験や運用時に観測されるシステムの

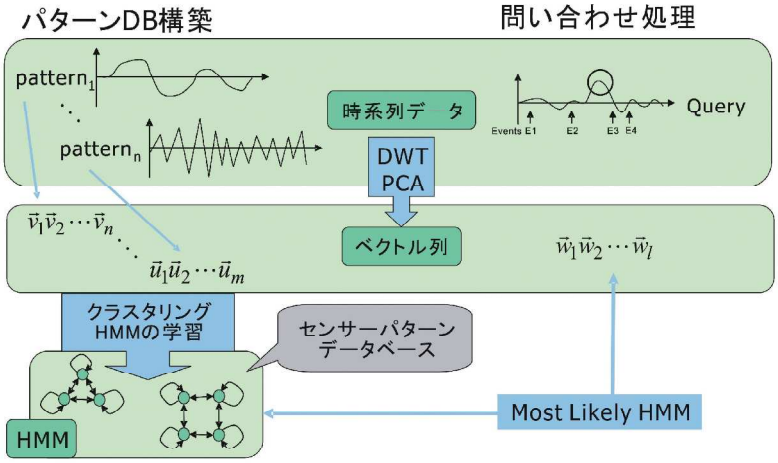

図５　センサデータによる検索

状態に対して、その状態が特異な状態であれば分析結果は報告書にまとめられ保存されることになる。そこで、それらの分析結果をまとめた報告書を観測されたセンサデータとあわせてシステムに蓄積していくことによって、観測されるセンサデータとシステムの状態に関するデータベースを構築することが可能になる。このデータベースを利用する方法としては、システム稼働時に観測されるセンサデータに対して、このデータベースに格納された類似センサデータを探し、システム状態のラベルを付与したり、過去の分析結果報告書にリンクを付けることによって、システム運用を補助するということが考えられる。MIMIRでは、センサデータを問い合わせに用いて、データベースを検索する方法関連情報を検索する技術についての研究を行った[７]。

この研究では、まず、システムの各状態で観測されるセンサデータをもとに、センサパターンデータベースを構築する。このデータベースでは、センサの状態を以下に述べる方法で構築された隠れマルコフモデルで表している。まず、各状態で出力されるセンサデータをウェーブレット変換し、特徴ベクトル列を生成する。そして、この特徴ベクトル列を学習データとして隠れマルコフモデルを構築する。隠れマルコフモデルを構築する場合に、そのグラフィカルな構造を決定することが大きな問題となるが、ここでは、ウェーブレット変換によって得られるベクトルをクラスタリングすることによって、状態を生成した。なお、センサパターンデータベースは実験時に得られたセンサデータを使って構築され、そこに含まれる各パターンには、実験の解析で得られる文書がリンクづけられている。

一方、稼働時にシステムから送られてくるセンサデータに対しては、センサパターンデータベースの構築と同様の手順で、ウェーブレット変換によって特徴ベクトル列を生成し、このベクトル列を問い合わせとして、センサパターンデータベースの検索を行う。ここでは、センサパターンデータベースに含まれる隠れマルコフモデルの中で、問い合わせにマッチするものが選択され、そのパターンに付随する文書が併せて得られる。図５は上記のパターンデータベース作成および問い合わせ処理の流れを表している。

5．おわりに

本稿では、システム開発において技術者を支援する情報共有システムMIMIRを紹介した。システム開発における情報共有の問題点のひとつ、技術者による情報の電子化支援にある。本稿では、この問題に対処するために、技術者にとって電子化の負担の少ない音声データやセンサデータを積極的に活用するためのMIMIRの試みに焦点をあてて、筆者が行っている研究の概要を述べた。紙面の都合で、評価実験の結果は省いたが、これらの技術を実用化するためには、その中で用いられるパターン処理の精度の向上が必要である。あわせて、パターン処理に多少の誤りがあっても問題のない、情報の利用方法についてもさらに検討を行うことが必要である。

参考文献

［1］ B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widow. "Models and issues in data stream systems," *SIGMOD'03,* pp.28-39, 2003.

［2］ S. Kahan, T. Pavlidis, and H. S. Baird. "On the recognition of printed characters of any font and size," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* 9(2):274-288, 1987.

［3］ E. S. Ristad and P. N. Yianilos. "Learning string-edit distance," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* 20(5):522-532, 1998.

［4］ F. Sebastiani. "Machine learning in automated text categorization," *ACM Computing Surveys,* 34(1):1-47, 2002.

［5］ S. Srinivasan and G. Petkovic. "Phonetic Confusion Matrix Based Spoken Document Retrieval". *23rd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval,* pp.81-87, 2000.

［6］ A. Takasu and K. Aihara. "DVHMM: Variable Length Text Recognition Error Model". *15th Intl. Conf. on Pattern Recognition,* pp.110-114, 2002.

［7］ A. Takasu and K. Aihara. "An Annotation Method for Sensor Data Streams based on Statistical Patterns". *24th International Conference on Database and Applications,* pp. 95-100, 2006.

［8］ M. Wechsler, E. Munteanu, and P. Schauble. "New Techniques for Open-Vocabulary Spoken Document Retrieval". *21st Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval,* pp.20-27, 1998.

［9］ 相澤, 大山, 高須, 安達. レコード同定問題に関する研究の課題と現状. **電子情報通信学会論文誌**, J88-D-I(2):576--589, 2005.