「高度メディア社会の生活情報技術」
平成12年度採択研究代表者

辻井　潤一

（東京大学大学院情報理工学系研究科　教授）

「情報のモビリティを高めるための基盤技術」

1.研究実施の概要

WWWに代表される情報ネットワーク中には、膨大な量のテキスト情報が蓄積されている。ただ、情報が物理的に蓄積されていることと、必要なときに必要な情報が参照できることとは別である。情報の氾濫は、氾濫の中の飢餓という奇妙な現象を引き起こしている。

この状況を解消し、テキストに埋もれた必要な情報を取り出し、それを理解しやすい形で提供するシステムを開発することが、本プロジェクトの目的である。

このためには、背景知識（オントロジー）を活用したテキスト理解という、非常に計算量が大きな処理を膨大なテキストに対して適用する必要がある。また、不特定多数の人間集団が持つ複数の背景知識を適切に使い分ける必要もある。とくに、情報要求者と情報提供者の背景知識の差異を、適切に解消することが不可欠である。

このようなことから、本プロジェクトでは、

（１）背景知識を活用するテキスト理解とテキスト・アーカイブ技術
（２）広域分散リソースを活用するソフトウェア技術
（３）テキストから背景知識を構築する技術
（４）背景知識を活用した情報提示の技術

を開発し、それらを統合することで、（A）処理時間・記憶容量の観点からも可能で、かつ、
（B）質的にも現在の検索技術をはるかに越える、情報活用システム(Information Exploitation System)を構築することを目指す。

これまでの２年半の研究においては、知識を活用した言語理解システム（Information Extraction System）、テキスト理解の結果を索引構造に反映するテキスト・アーカイブ技術、高速クローラー、分散リソースを活用するモーバイルエージェント技術、テキストからの背景知識獲得技術、テキスト情報を対話形式で提示する技術など、上記４つの分野の基盤的要素技術を開発してきた。プロジェクトの後半となる本年度以降は、これらを統合する技術の開発を行う。

2.研究実施内容

[研究目的]

膨大なテキスト集合の中から情報要求者に必要な情報を持ったテキスト箇所を同定し、背景知識と組み合わることで、個々の要求者に適合的な応答を返すことは、それほど容易ではない。言語処理と知識処理が持つ原理的むつかしさだけでなく、計算負荷の大きい２つの処理を、膨大なテキスト集合に対して実行するための技術的な課題を、まず解決しなければならない。このため、米国を中心に進められているQ/Aの研究では、通常のWeb検索手法をまず適用し、比較的、粗い処理で候補を絞った後に、言語処理、知識処理を起動するのが普通である。

しかし、この２段階での処理は、（１）本当は関連する情報を含むテキストを見落とすこと、また、（２）後段で起動する言語処理、知識処理も理解を経ない計算負荷の低い処理とならざるを得ないこと、から、必ずしも理想的なものではない。この種のアドホックな手法は、それ自体で閉じたシステムを構成し、Semantic WebやOntology技術といった将来進展するであろう、他の技術分野との整合性が非常に悪いものとなっている。

本プロジェクトの目的は、知識・情報・テキストには、それぞれの論理的構造があるとの立場から、極力アドホックな手法を避け、論理的に明確な構造を持ったテキスト情報管理システムを構築する。かつ、そのシステム構造の上に、確率・統計処理を載せることで、ユーザの背景知識に応じて柔軟に情報の検索・提示を行う手法を開発する。

[研究手法]

処理効率という技術的な困難を解消するために、本プロジェクトでは、

[A] 収集されたテキストに対して、あらかじめ計算負荷の大きい言語処理を実行し、その結果を索引構造に反映しておくことで、QA実行時の計算負荷を大きく軽減すること、

[B]概念階層による推論を高速化した論理型言語と専用の索引構造を用いることで、知識処理の計算負荷を軽減すること、

[C]知識資源がネットワーク中に分散した環境での最適な処理モデルを構築すること、

によって、２段階処理の欠陥を解消した、理論的にも整合性のあるQAモデルを構築する。

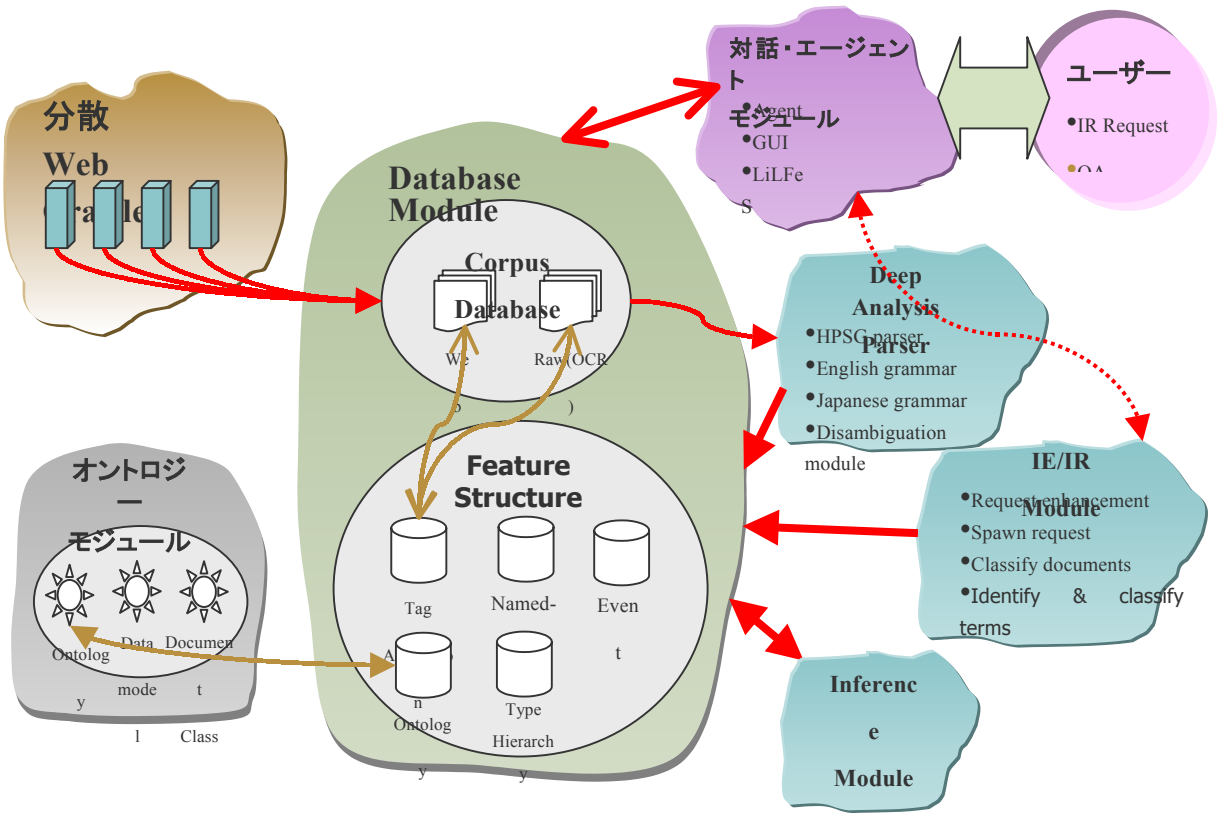
また、従来、深い知識処理と大量テキストの処理とが融合できなかったのは、処理効率だけではなく、大量テキストを処理できるだけの広範な知識が構築できなかったことも大きい。このことから、本プロジェクトでは、

[A]ゲノム生命科学など背景知識の重要性を認識し、その構築を行っている専門家グループと密接に協力し、大規模知識の構築を行うこと

[B]テキストから知識獲得、および、既存知識を有効に活用する異種オントロジー統合の技術を開発すること

[C]背景知識に基づく処理の有効性を情報抽出や検索だけで実証するのではなく、同一内容のテキストと対話の相互変換など、さまざまな局面で知識を活用する処理モデルを構築すること

を目指している。現在のプロジェクト概要を図１に示す。



［研究経過］

プロジェクト開始後、２年半を経過した現在の時点では、図中の各要素技術（大量テキストを取り扱う論理型言語、文法を用いた情報抽出、知識構造を用いたテキストの索引構造、高効率なWeb Crawlerなど）が完成している。また、大量の分野知識の構築では、生命科学の研究グループと協力することで、４０の意味クラスを設定し、この分類をもとに論文抄録（１００万語のMedline抄録）に意味タグを付与する作業が完了している。

3. 研究実施体制

言語処理グループ

① 研究分担グループ長：辻井　潤一（東京大学大学院情報理工学系研究科、教授）

② 研究項目：情報のモビリティを高めるための言語処理技術の研究および全体の統括

広域・分散ソフトウェアグループ

① 研究分担グループ長：米澤　明憲（東京大学大学院情報理工学系研究科、教授）

② 研究項目：言語処理のための大規模データ収集およびサービスシステム

オントロジーグループ

① 研究分担グループ長：中田　圭一（東京大学、教授）

② 研究項目：オントロジーのための記述枠組とオントロジー構築に関する研究

エージェント・対話グループ

① 研究分担グループ長：西田　豊明（東京大学大学院情報理工学系研究科、教授）

② 研究項目：情報提示とインターラクションに関する研究

4. 主な研究成果の発表（論文発表および特許出願）

（１）論文（原著論文）発表

○ Yuka Tateisi and Tomoko Ohta and Jing-Dong Kim and Masayoshi Tsuruoka and Jun'ichi Tsujii A Biomolecular Ontology as the Basis of Semantically-Annotated Corpora Proceedings of the Standards and Ontology for Functional Genomics (SOFG), November 2002

○ (?) Jin-Dong Kim and Jun'ichi Tsujii Copus-Based Approach to Biological Entity Recognition Proceedings of the Second Meeting of the Special Interest Group on Text Data Mining of ISMB 2002, August 2002

○ (?) Tomoko Ohta and Yuka Tateisi and Jin-Dong Kim and Jun'ichi Tsujii The GENIA Corpus: an Annotated Corpus in Molecular Biology Domain Proceedings of the 10th International Conference on Intelligent Systems for Molecular Biology (ISMB 2002) poster session, August 2002

○ Minoru Yoshida Extracting Attributes and Their Values from Web Pages Proceedings of the ACL 2002 Student Research Workshop, 72--77, 2002

○ Takashi Ninomiya and Takaki Makino and Jun'ichi Tsujii An Indexing Scheme for Typed Feature Structures Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), 1248-1252, 2002

○ Takashi Ninomiya and Yusuke Miyao and Jun'ichi Tsujii Lenient Default Unification for Robust Processing within Unification Based Grammar Formalisms Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), 744--750, 2002

○ Hideki Mima and Sophia Ananiadou and Goran Nenadic and Junichi Tsujii *A* Methodology for Terminology-based Knowledge Acquisition and Integration Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), 667—673, 2002

○ Jun'ichi Kazama and Takaki Makino and Yoshihiro Ohta and Jun'ichi Tsujii Tuning Support Vector Machines for Biomedical Named Entity Recognition Proceedings of the Natural Language Processing in the Biomedical Domain (ACL 2002), Philadelphia, PA, USA, July 2002

○ Tadayoshi Hara and Yusuke Miyao and Jun'ichi Tsujii Clustering for obtaining syntactic classes of words from automatically extracted LTAG grammars Proceedings of the sixth International Workshop on Tree Adjoining Grammars

and Related Frameworks (TAG+6), Venice, Italy, 227-233, May 2002

○ Naoki Yoshinaga and Yusuke Miyao and Jun'ichi Tsujii A formal proof of strong equivalence for a grammar conversion from LTAG to HPSG-style roceedings of the sixth International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6), 187--192, May 2002

○ Naoki Yoshinaga and Yusuke Miyao Grammar conversion from LTAG to HPSG WEB-SLS: the European Student Journal on Language and Speech, 2002

○ Toshiyuki Takahashi and Hong Soonsang and Kenjiro Taura and Akinori Yonezawa World Wide Web Crawler Proceedings of the 11th International World Wide Web Conference, 2002

○ Hideki Mima and Sophia Ananiadou and Goran Nenadic and Jun'ichi Tsujii *TIMS - A Workbench* for Ontology-based Knowledge Acquisition and Integration Proceedings of Natural Language Processing in Biomedical Applications (NLPBA 2002), 2002

○ Toyoaki Nishida Communicative Reality for Social Intelligence Design invited talk The IEEE International Workshop on Knowledge Media Networking (KMN'02), CRL keihanna, Kyoto, July 2002

○ Toyoaki Nishida *HAI in Community* Journal of Japanese Society for Artificial Intelligence, 17 (6) 665-671, 2002

○ Toyoaki Nishida Social Intelligence Design and Communicative Reality KES'2002 Sixth International Conference on Knowledge-Based Intelligent Information & Engineering Systems, Crema, Italy, 5-6, Sep. 2002

○ Daichi Shirayama and Keiichi Nakata Application of Text Analysis Techniques for Alignment of Ontological Categories Proc. of SSGRR-2003W, L'Aquila, Italy, 2003

○ Klaus Voss and Keiichi Nakata An Agent-Based Approach to Dynamic Ontology Construction Proc. of SSGRR-2003W, L'Aquila, Italy, 2003

（２）特許出願

H14年度特許出願件数：0件（研究期間累積件数：0件）