古典文献用 Digital Hermeneutics System の提案と実装

村井 源　　往住 彰文
東京工業大学　東京工業大学

従来のデジタルアーカイブは，テキストを電子化して検索などに用いるのが主であった．しかし，情報処理技術を活用することで，人間のテキスト解釈のプロセスまでをサポートすることが可能と考えられる．本論文では，テキスト解析に必要なアルゴリズムを柔軟に組み合わせ再帰的に利用可能な Digital Hermeneutics System のシステムデザインを行い，JAVA を用いて，サーバークライアントモデルのアプリケーションとして実装した．

Proposal of design and implementation of Digital Hermeneutics System for classic texts

Hajime Murai　　Akihumi Tokosumi
Tokyo Institute of Technology　Tokyo Institute of Technology

Former digital archives store electronic texts and those electronic texts are utilized for searching and so on only. But it seems to be possible to support a human process of interpreting text by utilizing information processing technology. This paper proposes a design of Digital Hermeneutics System which can combine various text processing algorithms and can utilize recursively. Digital Hermeneutics System was implemented as server client model application by using JAVA.

1．はじめに

古典文献に関する学問では，大量のテキスト資料等を必要とするため，旧来より文献の蓄積・保存・整理および検索などが大きな問題となってきた．これらの問題に対して近年急速に発達しつつある情報技術は有効であり，古典文献とその周辺分野でも種々の電子化プロジェクトが進行している．

旧来の多くの古典文献デジタルアーカイブでは，デジタル化の主目的が蓄積・検索・表示の効率化であった．またこれに合せて語の出現頻度や位置を表示するコンコルダンス機能や，テキストの単語と辞書の連動機能も見られる．この種のものでは WEB 上で公開されているギリシア語・ラテン語などの文献集の Perseus Project[1] (http://www.perseus.tufts.edu/) や，イスラム教関連の文献を集めた IIDL[2]，キリスト教関連文献集の CCEL(http://www.ccel.org/) などが良く知られている．テキスト中心のもの以外に，地図・写真・絵画などとのマルチメディア的なデータ関連機能を強化した製品などさまざまな種類のデジタルアーカイブも多数作られてきている[3].

これらの古典テキスト分野におけるデジタルアーカイブシステムは，人間がテキストを書き，読み，そして解釈して情報を得る行為の一部を，電子的なシステムに置き換えることで，従来にない機能や効率性を得ようとするものである．

まず，人間のテキスト利用プロセスを電子化することで以下のメリットが期待される．

1. テキストの電子化によるメリット
   - 保存の信頼性の向上，情報劣化の阻止
   - 蓄積・複写容易性
2. 電子化テキストの蓄積によるメリット
   - 情報整理の効率化
   - 検索コストの減少と大規模検索の実現
3. コンピュータの記号処理能力による大規模な自動的情報処理のメリット
   - 人手の場合では困難である，大規模な解析の低コストでの実現
   - アルゴリズムとして明示的・客観的にすることでの反証可能性の確保
   - 人手で見落としていた特徴の発見（テキストマイニング）

テキスト電子化のメリットに対して，人間が情報を伝達にテキストを利用するプロセスは次のように整理できる．

A) 情報をテキストで保存する（記録）
B) テキストのアーカイブ化（整理・保守，検索）
C) テキストを読み，理解する（理解・解釈）

双方を比較すると，従来の古典文献の電子化プロジェクトによって対象とされてきたのは，主に，テキスト電子化のメリットとその蓄積によるメリットを活用した B のアーカイブ化までであることが分かる．例外的に B の検索機能に関しては，コンピュータによる記号処理能力のメリットが活かされているといえよう．ただし，C の情報の理解（テキストの場合解釈）のプロセスについては，辞書とのマッチング（形態素のタグ付けなども含め）や単語のデジタルコン

コルダンス表示など基礎的なものを除きほとんど行われてこなかった．

このような現状に対し，テキストの電子化において，情報理解のプロセスまでをターゲットに含めることで，従来の電子書庫システム（Digital Archive System）を越えた電子解釈システム（Digital Hermeneutic System: 以下 DHS）の実現が可能になると考えられる．解釈のレベルにおいても，機械的な大規模情報処理を活用することで，効率的かつ従来にないテキストの読解が可能になると期待される．

一般的なテキストにおいても，解釈の困難なものは多数存在する．中でも，古典文献においては，非現代語の翻訳，時代的・文化的に距離のある社会の背景的知識の必要性，古代的な修辞法の理解など様々な原因によって解釈が困難になる場合が多い．

従来は，このような古典文献の解釈上の問題は，個々の学者が長年にわたって種々の文献から学び，修練によって匠のような研ぎ澄まされた読解力を身につけることで始めて克服することができた．そのため，専門的な知識を持たない一般人が古典文献を正しく解釈することは困難であった．専門家にとっても，解釈が可能な段階に到達するまでに長い年月を要するため，研究成果をあげることが他の理系分野などに比べて容易ではないという問題があった．

古典文献のような，テキストの解釈に膨大な知識と経験を必要とする分野においては，DHS のような情報処理システムをサポートツールとして用い，テキストにおける新しい発見や解釈のプロセスまでも含めてデジタル化することができれば，研究に大きなメリットが得られると期待される．また，専門的知識を持たない一般の愛好者が古典文献を読解する上でも，DHS は大きな助けになると考えられる．

2．DHS のシステムデザイン

2.1　古典テキストと引用関係

古典文献と呼ばれるようなテキストは，歴史を通じて大きな影響を及ぼしてきた．そのため，ある古典文献の周辺には，それに影響を受けた複数のテキストが存在している(間テキスト性[4])．また，古典テキストは後世に解釈され直される事で新たな意味を付与されることが少なからずある．このため，当該古典テキストのみではなく，古典テキストを解釈したテキストも合せて解釈することで，テキストが歴史の中でどのように捉えられてきたか，総体としてどのような概念形成に寄与してきたのか，を解析することが可能となる．

古典文献の中でも特に，宗教の正典等に代表される思想書は，たとえ話や比喩などの間接的で高度な解釈を必要とする表現に富んでいる．そのため，あるフレーズの意味を唯一の解釈に限定すること自体が困難である．唯一の解釈が特定困難な結果として，一つの文書に対して，多数のグループが作成した多数の解釈が並立する状況が築かれる．このような，各グループの教義を記した文書を集積することで，グループごとの教義の相違と関係性も解析できると考えられる．

たとえば，現代社会にも大きな影響を与えており，また多くのデジタルアーカイブプロジェクトの対象ともなってきた，キリスト教の聖書と関連文書は，正典である聖書の引用・解釈によって主に構成されている．このため，教義を記した文書中では古代より多数の聖書引用が行われる．また伝統的聖書解釈を重んじ正統的な教義であることを主張する意図からも教義文書間でも引用は頻繁に行われている(図 1)．

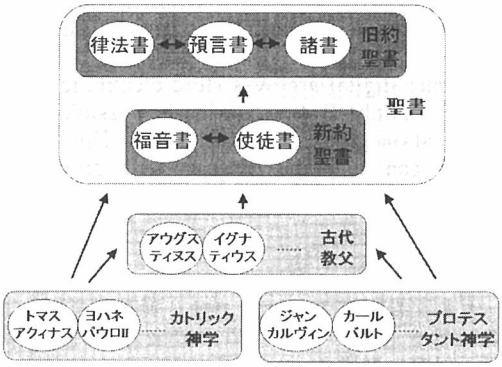

図 1　聖書における引用関係

このように，古典文献の解釈では，時間的・空間的な隔たりのある周辺テキストとの間の関係性によって新たな意味が派生しうる．そこで，周辺テキストとの間の参照情報を用いた古典文献デジタルアーカイブシステムを設計し，提案した(図 2)[5].

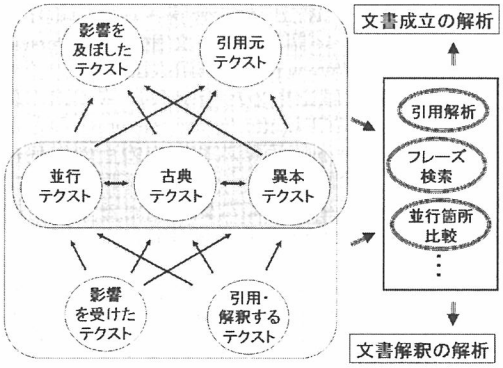

図 2　引用関係に基づくアーカイブシステム

2.2 解釈学に基づくシステムデザイン

古典テキスト解釈の際に考慮すべきテキストからの意味抽出のメカニズムとして，テキストとテキストの関係性である間テキスト性を挙げたが，それ以外にもさまざまメカニズムがある．たとえば，テキスト解釈において基本的な理論である解釈学[6]においては，「解釈学的循環」や「地平の融合」などの概念が重要なものと考えられている．「解釈学的循環」は解釈の結果としての知識構造が，さらなる解釈の背景知識として再帰的に活用される循環構造である．「解釈学的循環」を DHS で実現するには，解析の出力を再帰的に解析の入力として利用可能なシステム構成が必要である．「地平の融合」は，異なる知識構造の解釈を通じた融合とも言える．「地平の融合」を実現するには，背景的知識に相当する入力の相違を反映した出力の相違と，これらの知識構造の「解釈学的循環」による接近，収束を実現するシステムが求められる．

このため，テキストの自動的解釈には，従来の様々なテキスト処理技術を活用しつつ，「解釈学的循環」や「地平の融合」のメカニズムにも即した，システムデザインが必要である．具体的には，種々の解析アルゴリズムを柔軟にモジュールとして組み込み可能であり，なおかつ，各モジュールで用いられる入出力データ形式が，他モジュールでも再帰的に利用可能である必要がある．また，これらの入出力形式が従来の知識構造の表現にも対応可能なことが望ましい．

2.3 テキスト解析に用いられる処理

テキスト処理には目的に応じて種々の解析手法が存在するが，本論文ではテキスト解釈に関わる主な分野に対象を絞り，テキストや言語情報を基に解析する基礎的解析手法と，他の解析手法の結果を複合する複合的解析手法に分ける．さらに基礎的解析手法を，主に用いるデータで，言語知識・テキスト内情報・テキスト間情報に基づく三種類に分類する．

まず言語知識に基づく基礎的解析手法としては，電子化辞書やオントロジー，形態素解析・構文解析，修辞解析が挙げられる．電子化辞書やオントロジーと呼ばれる分野は単語の意味を，機械可読の状態で定義し，機械による意味処理を可能にするため，研究されてきている．形態素解析や構文解析では，文章の中の単語を品詞情報なども合わせて特定し，それらの構文上の関係性を特定するものである．形態素解析には，電子化辞書やオントロジーの情報が必要となる．また，修辞解析は各言語・時代ごとの修辞法の規則や，構文解析の結果などに基づいて，意味的関係を解析する手法である．テキストの修辞構造を XML 形式などで記述する手法[7]が開発されているが，自動的な修辞構造の認識は未だ困難なのが現状である．これらの手法はみな，言語・文法的な知識に基づいて処理が行われる点が共通している．ただし，修辞解析には文法的な知識のみではなく，テキスト内の他の部分との関係性の情報が必要となる．

テキスト内情報に基づく基礎的解析手法としては，種々のキーワード抽出・テーマ解析や，コンコルダンス解析などを挙げることができる．テキストの主要な内容・テーマを一つの単語，あるいは複数の単語群として表現し，これを自動的にテキストから抽出する手法はキーワード抽出と呼ばれる．あるテキストの部分の特徴を示すキーワードは，次のような条件のいずれかを満たすと考えられている[8]．

- 頻出する語
- 当該部分のみに頻出する語
- 他の部分に比べて当該部分に頻出する語
- 全体の頻度分布から考えて当該部分では偏った出現の仕方をする語

コンコルダンス解析は，テキスト内で出現する各語彙の使用法を列挙し，そこから語彙のニュアンスを特定する手法である．テキスト内の語彙使用の列挙までが自動的に行われる場合が多い．キーワード抽出も，コンコルダンス解析もテキスト内の情報に基づいているという点が共通している．

テキスト間情報に基づく基礎的解析手法としては，科学論文の分類に頻繁に用いられる引用解析が挙げられる．引用解析は，テキストの分類以外にも，周辺テキストにおける特定のテキストの価値や意味合いを判定するのに用いることができ，ネットワーク上のテキストにおける重要度の判定や，テキストコーパス内での役割の判定などが行われている．

これらの基礎的な解析手法による結果を合わせた，高次レベルの複合的な解析手法としては，計量文体学に基づく文体解析や，キーワードの変化などを用いるテキストの構造解析，さらにテキストの内容自体を簡潔に抽出するための種々の要約手法などを挙げることができよう．

上記の，主な解析手法の関係性を図 3に示す．

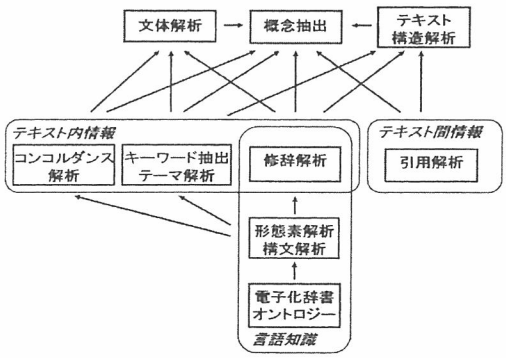

図 3 テキスト解析手法の関係

DHS においては，これらの種々のテキスト解析手法の入出力が，再帰的に利用可能なデータ形式を設定する必要がある．

2.4 テキスト解析に用いられるデータ

2.3 で挙げたような，従来の主なテキスト解釈に用いられる技術の入出力の形態を調べていくと，これらは要素の単体，要素のリスト，要素のペア，そしてネットワークであることが分かる(表 1，表 2)．これらの諸形態をオブジェクトとし，形態間の変換を各言語処理モジュールによって実現するオブジェクト指向のシステムを構築する．要素の単体，要素のリスト，要素のペア，そしてネットワークの相互の形式の変換は基本的には図 4のように考えることができる．

入出力にテキスト解釈で用いられる一般的なデータ形式をオブジェクト化して採用する．そして，オブジェクトの再帰的な処理を，ユーザーの用途に応じて追加可能なモジュールによる相互変換という形で実現するする．オブジェクト化とモジュール化によって，解釈学的メカニズムを実現し，かつ拡張性の高い汎用的テキスト解釈システムを構築可能であると考えられる．

表 1　テキスト解釈の手法と入力データ形態

	出力テキスト要素	形態	利用法
電子化辞書・オントロジー	文	単体	意味定義
	単語	リスト	同義語・類義語
	単語	ネットワーク	関連単語関係
形態素解析・構文解析	単語	リスト	派形単語リスト
	単語	ネットワーク	単語間の係り受け
	文節	ネットワーク	文節間の係り受け
修辞解析	文節／文	ネットワーク	修辞構造
	単語	単体	使用修辞法の特定
キーワード抽出・テーマ解析	単語	リスト	キーワード
	単語	ネットワーク	キーワードの関係性
コンコルダンス解析	節(部分テキスト)	リスト	単語使用の文脈
	単語	リスト	近傍単語のリスト
	単語	ネットワーク	近傍単語の関係性
引用解析	テキスト(全体／部分)	リスト	引用・被引用リスト
	テキスト(全体／部分)	ネットワーク	引用テキストの関係

表 2　テキスト解釈の手法と出力データ形態

	入力テキスト要素	形態	利用法
電子化辞書・オントロジー	単語	単体	
形態素解析・構文解析	文	単体	
修辞解析	テキスト(全体／部分)	単体	修辞構造
	節(部分テキスト)	単体	使用修辞法の特定
キーワード抽出・テーマ解析	テキスト(全体／部分)	単体／リスト	
	節(部分テキスト)	単体／リスト	
コンコルダンス解析	テキスト(全体／部分)	単体／リスト	
	単語	単体／リスト	
引用解析	テキスト(全体／部分)	単体	

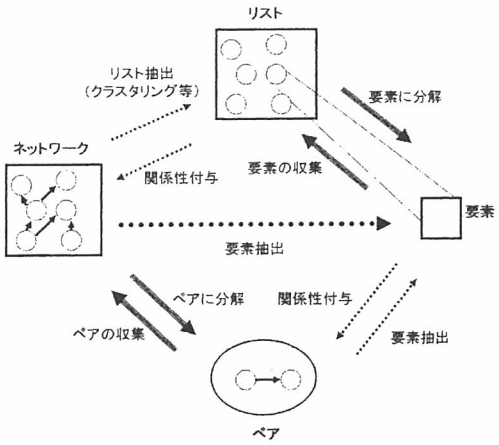

図 4　データ諸形式の関係性

2.5 テキスト解析の流れ

このようなデータ形式のオブジェクト指向，モジュール間の連携の構造が，実際のテキスト意味解析に有効である例として，コンコルダンス解析の場合を図 5に示す．

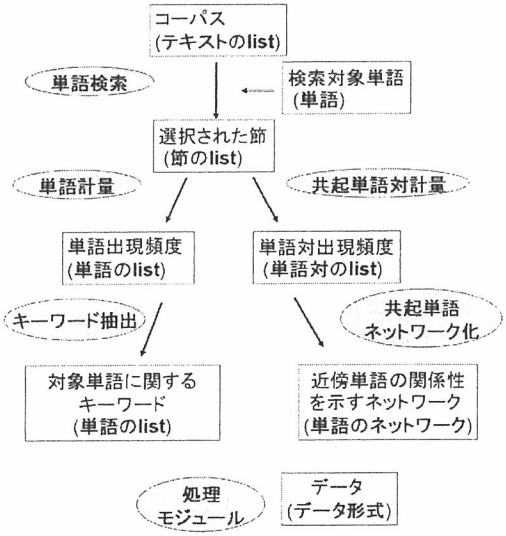

図 5　コンコルダンス解析での処理例

このように，オブジェクト化された各形式のデータを複数のモジュールを用いて変換することで，テキスト解釈のための計量的な基礎データを出力することが可能である．

DHS は，時空間的に隔たりのある周辺テキストとの関係性から発生する意味の解析機能に合わせて，人間の一般的なテキストの解釈を計量的に客観化するためのサポートツールを目指す．

3. DHSの実装

2節で提案したシステムデザインに基づき，JAVAとTomcatを用いたサーバー・クライアント型のアプリケーションとしてDHSの実装を行った．電子化対象としてはキリスト教の聖書とその周辺文書をターゲットとし，7種の翻訳(言語数は4)の聖書（表3）と，聖書引用箇所のデータベース(MySQL使用，表4)約40万件をデータとしている．

表3　DHSにアーカイブ化したテキスト一覧

旧約聖書	Septuaginta (70人訳聖書)	聖書ギリシア語
	Nova Vulgata	ラテン語
	King James	英語
	New American Bible	英語
	New Revised Standard Version	英語
	新共同訳	日本語
旧約聖書続編	Nova Vulgata	ラテン語
	New American Bible	英語
	新共同訳	日本語
新約聖書	Nestle-Aland26版	聖書ギリシア語
	Nova Vulgata	ラテン語
	King James	英語
	New American Bible	英語
	New Revised Standard Version	英語
	新共同訳	日本語
	フランシスコ会訳	日本語

表4　DHSに取り込んだ聖書引用データ

著者	著作数	引用数(節単位)
アウグスティヌス	43	37832
トマス・アクィナス	32	24118
ジャン・カルヴァン	47	113515
カール・バルト	113	185915
ヨハネ・パウロ二世	1939	57372

本論文で実装した，解析機能を実行する各モジュールおよびJAVAのクラスと，2.3で示したテキスト関連の情報処理機構との対応関係を表5に示す．修辞解析機能に関しては現状ではその自動的実現が困難であるため，現バージョンのシステムにおいては未実装となっている．複合的な解析機能は，基本的には複数の基礎的解析モジュールを組み合わせて実現する．複合的な解析に用いられる入出力データの多くは，多数の要素とその関係性で記述されるネットワークである．このため複合的解析機能は，いくつかのモジュールの結果を，ネットワーク解析モジュールで解析することで実現可能な場合が多い．

表5　解析機能とモジュール

電子化辞書・オントロジー	jp.ac.titech.lkr.library.dictionary のクラス群
形態素解析・構文解析	jp.ac.titech.lkr.library.dictionary のクラス群
キーワード抽出・テーマ解析	単語計量モジュール＋ ネットワーク解析モジュール
コンコルダンス解析	多言語検索モジュール＋ 単語計量モジュール＋ ネットワーク解析モジュール
引用解析	引用解析モジュール＋ ネットワーク解析モジュール

解析に用いるオブジェクトは要素の単体と，ペア，リスト，ネットワークである．これらのうちリスト（ペアのリストも含む）はcsv形式で，ネットワークはdot形式（Graphviz用[9]）およびnet形式（Pajek用[10]）で読み出し・保存が可能である．

オブジェクト型の解析モジュールによる変換である，現バージョンの関係を図6に示す．引用閲覧・解析モジュール，多言語検索モジュール，単語計量モジュール，ネットワーク解析モジュールの各モジュールに対し，単語・節（部分テキスト）・テキストのそれぞれが，単体（element）・リスト・ペア・ネットワークの形態で，入出力として用いられるようになっている．電子化辞書および形態素解析機能はモジュールではなくJAVAのクラス・関数群として実装しており，各モジュールにおいて必要な場合に呼び出される形式をとっている．

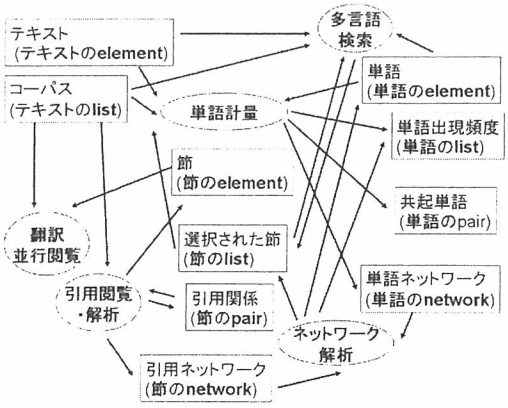

図6　モジュールとオブジェクト間の関係

現バージョンにおける，単語計量モジュールのスクリーンショットを図7に示す．

本論文が提案するDHSのシステムとデータを用い，実際の古典文献の意味解釈を行った成果としては，キリスト教における新約聖書の解釈体系の時代的・宗派的相違を引用解析から計量的に求めた研究[11]や，やはり新約聖書四福音

書間の記事の相違から，正典編纂者の聖書解釈を抽出する研究[12]が行われている．

また，現バージョンでは未実装であるが，複合的解析の結果のデータを各解析モジュールのパラメータとして用いるように改変することで，「解釈学的循環」に対応するシステムを実現可能である．たとえば，キーワード抽出の出力結果をオントロジーモジュールのデータとして保存可能にすることで，特定コーパス内のキーワードのオントロジー構造を解析し，各オントロジー構造の差異を反映しつつ他コーパスのテーマを解析することなどが可能になる．この例では，解析に用いるオントロジー，つまり知識構造の元となるテキストコーパスの差異が，新しいテキストの解釈に影響を及ぼすメカニズムを再現可能である．人間の場合で考えると，特定のテキスト集合で表される背景知識構造を用いて，あるテキストを解釈するとどのような理解を得られるか，という問題に対応している．

また，「解釈学的循環」に対応する処理の再帰的な繰り返しによる結果の収束や，複数種類の背景知識を相互に循環させた場合を解析することで，解釈における「地平の融合」の挙動を把握することが可能となる．

図 7　単語計量モジュールの画面

4．結論と今後の課題

従来の電子図書システムに合わせて，テキストの意味解釈のサポートも行う，汎用的で拡張性の高い Digital Hermeneutics System (DHS) のデザインを提案した．DHS はアーカイブ化されたテキストの間テキスト性をデータ化しており，時間的な解釈の変遷を計量的に知ることが出来る．またキリスト教の聖書とその関連データを用いて JAVA による実装を行った．本システムは現在 http://www.cathnet.org/での公開に向けて準備中である．

現バージョンは解析モジュールが基本的なものに限られているが，将来的にはオープンソースウェアとして公開し，ユーザーが必要に応じて独自の解析モジュールを追加できるような形で配布予定である．

参考文献

[1] Crane, G.,Wulfman, C.E., Cerrato, L.M., Mahoney, A., Milbank, T.L., Mimno, D., Rydberg-Cox, J. A., Smith, D. A. and York, C.: Towards a Cultural Heritage Digital Library, 3rd ACM+IEEE Joint Conference on Digital Libraries, 2003.

[2] Library, I. I.D.: http://www.iidl.net/.

[3] JBible, 日本コンピュータ聖書研究会, http://jcbr.gospeljapan.com/

[4] Julia Kristeva, *Le Texte du roman*, 1970, 『テクストとしての小説』谷口勇訳，国文社，1985.

[5] 村井源，三宅真紀，赤間啓之，中川正宣，徃住彰文: テクスト間参照情報を考慮した古典文献デジタルアーカイブ, IPSJ SIG Computers and the Humanities Symposium 2005, Vol. 2005, No. 21, pp. 21-25, 2005.

[6] Hans Georg Gadamer, Wahrheit und Methode: Grundzuge einer philosophischen Hermeneutik, 1960, 『真理と方法』, 轡田 收 訳, 法政大学出版局, 1986.

[7] Mann, W. C. & Thompson, S. A., "Rhetorical Structure Theory: Towards a functional theory of text organization", Text, vol. 8, no. 3, pp. 243-281, 1988.

[8] K. Kageura and B. Umino, "Methods of automatic term recognition: a review", Terminology, vol. 3, no. 2, pp. 259-289, 1996.

[9] http://www.graphviz.org/.

[10] http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

[11] 村井 源, 徃住彰文, "Co-citation Network による宗教思想文書の解析", 人工知能学会論文誌, Vol.21, No.6, pp. 473-481, 2006.

[12] 村井 源, 徃住彰文, "正典テキスト群から編集的中心メッセージを抽出するネットワーク解析法", 情報知識学会誌, Vol. 16, No. 3, pp.149-163, 2007.