情報処理学会論文誌 Vol.62 No.10 1748–1761 (Oct. 2021)

LMBC：スポーツ史における史的書簡管理システムの設計と試作

伊藤 秀昭[1,a)] 來田 享子[2]

受付日 2021年1月18日, 採録日 2021年7月7日

概要：本稿では，スポーツ史における史料の一種である書簡の分析を支援するために開発を進めている書簡管理システム LMBC の設計と，プロトタイプシステム開発について述べる．LMBC の処理対象は，第5代 IOC 会長ブランデージが保存していた書簡である．本システムの主たる機能は，書簡のデータベースへの定義機能，検索機能，書簡の交換と差出日に基づき関連する書簡の表示機能，および対象の関連付け機能である．一方，書簡では人物名や組織名などの固有表現が検索や分析において重要な役割を果している．固有表現を検索に利用するだけではなく，関連のある対象を求める機能が要求されている．本システムは，固有表現の共起関係から共起ネットワークを構築して，2つの固有表現を関連付ける固有表現の集まりを求める．関連付けの実験では精度と再現率を求めた．

キーワード：書簡管理システム，書簡分析，書簡検索，書簡関連付け，固有表現，関固有表現間の関連

LMBC: Design and a Prototype System of a Letter Management System in Sports History

Hideaki Ito[1,a)] Kyoko Raita[2]

Received: January 18, 2021, Accepted: July 7, 2021

Abstract: This paper describes the design and the prototype system of LMBC, to analysis letters in sports history, which are collected by Avery Brundage, the fifth president of I.O.C. The system is developed for organizing, retrieving, and relating the archived letters. The main functions of this system are describing letters in a database, showing letters by relating them based on correspondence and dates. On the other hand, entities play an important role in the analysis of letters, which are the names of persons and organizations, etc., called named entities. In addition, to find relevant entities to certain named entities is required. A co-occurrence network of named entities is constructed, and the relatedness between two named entities is sought. Recall and precision of retrieval results are measured.

Keywords: letter management system, letter analysis, letter retrieval, relating letters, named entity, relatedness of named entities

1. はじめに

近年，スポーツ史やオリンピック史を研究するために文書の集まりからなるデジタルアーカイブを構築して，情報技術を活用したアーカイブの内容の分析に対する要求が高まっている [1]．スポーツ・オリンピック史における歴史的資料（史料）の1つにアベリーブランデージコレクションがある [2]．このコレクションは，第5代 IOC（International Olympic Committee）会長 Avery Brundage（アベリー・ブランデージ）が活動初期より保存，収集した史料であり，現在イリノイ大学において一部が公開されている [3], [4]．コレクションには，文書の一種として書簡が含まれており，重要な史料の一種となっている．第2次世界大戦前後には書簡が主な通信手段であった．ただし，多くの文書はテキスト化されておらず，マイクロフィルムに保存されている．

[1] 中京大学工学部
School of Engineering, Chukyo University, Toyota, Aichi 470–0393, Japan
[2] 中京大学スポーツ科学部
School of Health and Sport Science, Chukyo University, Toyota, Aichi 470–0393, Japan
[a)] itoh@sist.chukyo-u.ac.jp

書簡の集まりが計算機可読となり，提示されたり，検索されたりすることは有意義である．

書簡の管理や検索を支援するシステムに対する要求に応えるために，我々は，LMBC（Latter Management System for Avery Brundage Collection）と呼ぶ，書簡管理システムの開発を進めてきた [5], [6], [7]．本研究は，ブランデージコレクションに含まれる書簡の解読や分析を支援するためのシステムを設計・開発することを目的としている．書簡分析では，効率的に書簡のテキストを参照して，関連する書簡を探すことが要求される．また書簡は，多くの人物や組織などの分野固有の対象やキーワードを用いて記されている．関連のある対象やキーワードを見つけて，互いに関連付けることによって書簡は分析される．特定できる対象や事物の名前は，固有表現と呼ばれている [8], [9]．

史料の利活用と，図書・書簡の組織的な構造化と公開を目的に種々なデジタルアーカイブが構築されてきた [10]．また，いくつかのデジタルライブラリでは，インターネットで文書や書簡を公開している．たとえば，フランス国立図書館のデジタルライブラリ Gallica [11] や，早稲田大学図書館古典籍総合データベース [12] では，図書資料の一部として書簡が公開されている．ケンブリッジ大学のデジタルライブラリでは，書簡の画像とテキストデータとが公開されている [13]．このライブラリでは，書簡や文書に対するメタデータや，IIIF を用いた画像の注釈やダウンロードが利用可能である．また，一部の文書には注釈や要約が付与されている．さらに，リード大学の Letters Database は，書簡に特化したデジタルライブラリである [14]．

一方，書簡は通信手段であり，現在の通信手段の 1 つは E メールである．E メールのための管理システムや検索システムが開発されている．文献 [15] では，E メールの管理は，個人情報管理の応用と見なされる．また，E メール群を分類，可視化および要約するシステムが開発されている [16], [17]．さらに，E メールの集まりから，トピックや頻出単語の時間に沿った変化を，表示する方法が開発されている [18]．一方，文献 [19] は，E メールの送受信を可視化することで，コミュニティを発見する方法の有効性を示している．E メールの処理システムとは扱う量の点では大きく異なるが，LMBC は時間に沿った書簡の表示や，個々の書簡の交換を示す機能を備える．ただし，E メール処理では，大量性のために詳細よりもむしろ概要を示す要求に応える必要がある．しかし書簡処理では，個々の書簡や書簡の交換を示す必要がある．

対象や対象間の関連を理解することによって書簡を分析する．ある 1 つの対象に関連する対象を求める方法に，グラフ構造を元にしてコミュニティを求める方法が提案されている [20]．文献 [21] は，クリークに基づく固有表現のコミュニティを発見する方法を提案している．コミュニティは，ある 1 つのノードに関連するノードの集まりである．文献 [22] では 2 つの対象間を結合するコミュニティを，電気理論に基づいて求めている．文献 [23], [24] では，それぞれグラフのノード，およびグラフの部分グラフ間の関連を求める方法を提案している．また文献 [25] は連想に基づき 2 つの語彙間の関連を計算している．さらに文献 [26] は，2 つの対象間の関連を知識グラフより求めている．LMBC は，ある 2 つの固有表現間に関連する固有表現を，共起関係に基づき求める．まず，固有表現が同じ書簡に出現する共起関係に基づいて，固有表現共起ネットワークと呼ぶネットワークを構成する．固有表現共起ネットワークにおける個々の固有表現の媒介中心性とクラスタ係数に基づいて，2 つの固有表現間の関連を求めた．媒介中心性とクラスタ係数は，ネットワーク構造におけるノードの特徴である [27]．

本稿は，以下のように構成されている．2 章では，LMBC 開発の動機について述べる．システムの設計方針や考慮した点は 3 章に示す．本システムのシステム構成は 4 章に示す．5 章には，書簡を定義する XML 表現と，固有表現のタグ付けを示す．検索結果の表示方法は 6 章に示す．7 章では，固有表現共起関係ネットワークに基づき，関連すると見なされる固有表現の集まりを得る実験について述べる．8 章にシステムに関して考察する．9 章に，本研究のまとめについて述べる．

2. システム開発の動機

一般に書簡データベースにおいて書簡を記述する項目には，書誌学的な立場から，枚数，発信地，受信地，補記，料紙，筆記具，切手，郵便料，消印などがある [14]．また，書簡の保存と修復のためのデータベース開発が進められており，料紙の特徴や素材が記述されている [28]．

一方，書簡の記述内容の分析が必要になったとする．書簡テキストを提供するデジタルライブラリ，たとえばケンブリッジ大学のデジタルライブラリ [13] に格納された書簡を内容分析するとき，差出人，受取人，発信地などはデータベースに記述されているので，それらは書誌的な分析に利用できる．しかし本文に出現する対象や別表記された対象を抽出する必要がある．

本研究において対象としたブランデージコレクションは，文書をマイクロフィルムとして保存する画像の集まりである[*1]．我々は，書簡の公開や保存・修復のためのデータ保存という立場ではなく，書簡の調査を支援するという立場からシステム開発を進めた．

ブランデージコレクションはスポーツ・オリンピック史において貴重な史料であり，一部が公開されている [3], [4][*2]．

[*1] 我々が対象とした書簡の集まりでは，書簡が封入された封筒や切手の画像は含まれていない．

[*2] いくつかの機関がコレクションを有しており，中京大学はコレクションを有する機関の 1 つである．

文献 [2] はコレクションのカタログ情報である．コレクションには，書簡，議事録，種々のレポート，IOC 関連記事，テレックスなどが含まれている．特に IOC 委員との意見交換が残されている書簡の集まりは，オリンピック史における研究対象となっている．

書簡の記述内容を分析する理由は，次のとおりである．

- ブランデージおよびオリンピック研究の 1 次史料であること
- 公的記録ではない私的領域での意見交換が記されていたり，人間関係が示されていたりしていること
- ある特定の目的のために，相手を特定して伝えていること
- これまで知られていなかった事実や，新たな歴史研究のテーマを発見できる可能性があること

ブランデージコレクションにおける書簡の分析は，次のような方法で進められている．利用者は，個々の画像を読みながら書簡を探したり，アーキビストに依頼して書簡を入手したりする．記述内容と時間関係から関連する書簡を探す．このような調査は効率的ではなく，また目録と必要な情報とが相違することが考えられる．たとえば，ブランデージコレクションの書簡を 1 次史料として利用した研究に文献 [29] がある．先に述べたように画像を参照しながら，ブランデージと永井松三（IOC 委員）との往復書簡を探すという作業によって，研究が進められた．このような研究を支援するため，LMBC が想定する利用者は，ブランデージコレクションに興味のある研究者や利用者である．

一般に史料分析では，書簡の緻密な精読により内容が分析される [30]．また固有表現として表される対象や，対象間の関係を調べる．たとえば，個人と組織，スポーツと大会，組織間の関係など．精読を支援するための情報システムを開発することによって，情報検索に関わる負担の軽減を図り，直感的にとらえることが難しいような対象間のつながりを見つけることが支援できると考えられる．

LMBC を開発する動機は，次のようなことである．

- 計算機可読とする XML 形式によって書簡を定義して，書簡群を管理するシステムを開発すること．書簡を表現するためのデータ構造を設計する．
- 書簡の検索を容易とする検索機能を提供すること．書簡の検索では，書簡の差出人や受取人を指定できる．
- 書簡の分析に必要な基本的機能を備えるシステムを開発すること．書簡はむろん文書であり，基本的機能として語彙の出現，語彙の共起関係などを，求める必要がある．書簡には種々の固有表現，スポーツやオリンピック分野での固有な語彙，それらの略記号などが出現する．それらを定義して，互いの関連を求める．

書簡は意見交換のための通信手段であり，個々の目的に基づき記述された文書である．書簡表示や分析支援のために，書簡の交換の様子や，時間順に交換された書簡本文に含まれる語彙を表示する．これは，書簡を交換している人々のネットワークや話題を明らかにして，関連する書簡や事項を，時間や空間的な距離に影響されることなく，書簡を有機的かつ組織的に関連付けるためである．このとき，書簡の形式が定型的であり，差出人，受取人および差出日が記述されていることが利用できる．たとえば，書簡には往復書簡であることは直接示されていない．しかし差出人や受取人，差出日などから書簡交換の様子が提示できれば，往復書簡であろう書簡群を見つけるためには有用である．

また，固有表現やキーワードが重要な役割を果たしており，2 つの固有表現やキーワード間に直接的な関連はないが，他の固有表現を介して関連があるのならばそれらを関連付ける．たとえば，“日本人” と “オリンピックムーブメント” という用語は同一の書簡に現れないので，これらが直接関連付けられることはない．しかし “国際オリンピック委員会（IOC）” を介して 2 つの概念は関連付けられる．

3. LMBC の設計

LMBC 開発の目的は，次の 2 つである．1 つは，書簡間の関係の理解を助けるために，ある書簡に関連する他の書簡の調査を支援することである．書簡の構造的な特徴は差出人，受取人および日付が記載されていることであり，これらの要素に基づいて関連する書簡の調査を容易にする．また，書簡間の関係の理解を助けるために，時間的・内容的に関連する書簡を提示する．書簡交換の様子や，時間軸に沿った書簡に出現する語彙の変化を示す．

もう 1 つの目的は，書簡に出現する固有表現やキーワードの関連を分析するために，固有表現を関連付けることである．1 つの書簡に出現する固有表現は互いに関連するので，固有表現の共起関係に基づき固有表現を関連付ける．このとき固有表現の共起関係を示すだけではなく，2 つの固有表現を関連付けることを目的とする．

また，システム開発に際して要求された事項は，次のようなことである．

- 書簡の分析支援に供する基本的な機能を有すること．書簡に出現する語彙の計数，語彙の共起関係を求める機能に加えて，差出日の順序に従い書簡を整列する機能，書簡の交換などの表示機能が要求される．
- 固有表現を定義して，表示や関連分析に利用すること．固有表現には人名や場所，機関名だけではなく，分野固有の組織名，大会名，会議名などとそれらの同義語がある．
- 書簡の特徴を反映した書簡を操作するインタフェースを備えること．利用者は，システム開発者だけではなく，スポーツ史に興味がある利用者が想定されている．たとえば，複雑な検索式を記述するというよりむしろ，書簡の構成要素を反映した検索を容易とする．

本システムの設計方針は，次のとおりである．

- 書簡の記述および検索のために書簡の構成要素を反映すること．書簡の構成要素には，差出人，受取人，固有表現などがある．これらの書簡を特徴付ける要素に基づき，XML 形式によって記述する．
- 固有表現を定義して検索に利用すること．現時点においては，固有表現はシステム開発者や利用者が定義する．
- 固有表現間の関連を求めるために，固有表現共起ネットワークを構成して，固有表現共起ネットワークの基本的なパラメタを利用して関連を求めること．
- 書簡の分析機能は要求に応じて構築すること．分析に必要な機能は，要求に応じて段階的に実装する．
- ケーススタディを通じて，システムの拡張を容易とすること．

一方，書簡を記述するデータ項目は，次のように設定することとした．書簡に記入された情報を構造的に記述するための項目や，書簡の交換を E メールの交換と見なして書簡の交換を記述するための項目を設ける．設けたデータ項目は，差出人と受取人の氏名，所属，住所，日付（差出日），主題，本文と本文を構成する段落，コピーの受取人，書簡の画像とテキストである．ただし，現時点では紙面に記載された手書きのメモは扱っていない．

テキスト分析システムとしての機能について，次のように考えている．現時点においては汎用的な分析機能というよりむしろ，システムは分析に要する書簡に出現する語彙や，語彙の共起関係を表示するための基本的な機能を備える．なお，基本機能として必要な出現する語彙を用いた検索や，語彙の共起関係などを求めて表示する機能の詳細は，[5], [7] を参照されたい．

システム構築に関して具体的に考慮した点は，以下のようなことである．書簡は，差出人と受取人の間で交換されるので，書簡の差出日（時系順）に基づき整列されて，参照されることが多い．LMBC では書簡の基本要素（受取人，差出人，差出日，本文）を用いた検索機能を備える．また，特定の期間に交換された書簡を調査することが多いので，検索結果の表示のために，時間の単位と表示枠の大きさを設定可能とする．指定された期間における書簡の分布を表示するためである．さらに，検索結果を参照しながら，書簡の画像を参照可能とする．

また，書簡にはスポーツ・オリンピック史に関連する固有表現が出現する．固有表現を自動的に文書から抽出するには，領域依存の辞書や知識が必要である．既存の固有表現抽出ツール [8], [9] から得られる抽出結果を，ただちに利用することは困難であると考えられる．ただし，固有表現抽出ツールを用いればすべてではないが特定の組織や人名などの固有表現を抽出できる．しかし書簡では正式な名称だけではなく，組織や人物などが異なる表記で表されたり，略記されたりすることが多い．既存の固有表現抽出ツールは同義語の処理を備えていないので，当面は利用者や開発者が固有表現やキーワード[*3]を同義語とともに辞書に定義することとした．

さらに，本システムでは，固有表現間のつながりを求めるために，固有表現共起ネットワークを構成した．書簡による意見交換において，1 つの書簡はある特定の話題に関して述べるので，関連のない対象が記されることは少ないであろう．したがって，共起する固有表現が互いに関連すると見なすことは適切である．また，リンク付けられた固有表現は，ある書簡で同時に言及されたことを示すので，リンクの解釈は容易である．

4. システム構成

本節では，LMBC のソフトウェア構成について述べる．LMBC は，次のモジュールからなる．それらは，(1) 書簡データベースを構築する書簡データベース管理モジュール，(2) 固有表現を定義して，本文に出現する固有表現をタグ付けする固有表現定義モジュール，(3) 問合せを処理する問合せ処理モジュール，(4) 問合せ処理のために必要な内部データを作成するデータ構成モジュール，および (5) ユーザとのインタフェースとなり，種々のモジュールの機能を呼び出すユーザインタフェースモジュールである．LMBC の概念的な構造を**図 1** に示す．

書簡は受取人や差出人ごとに整理され，ボックスに収め

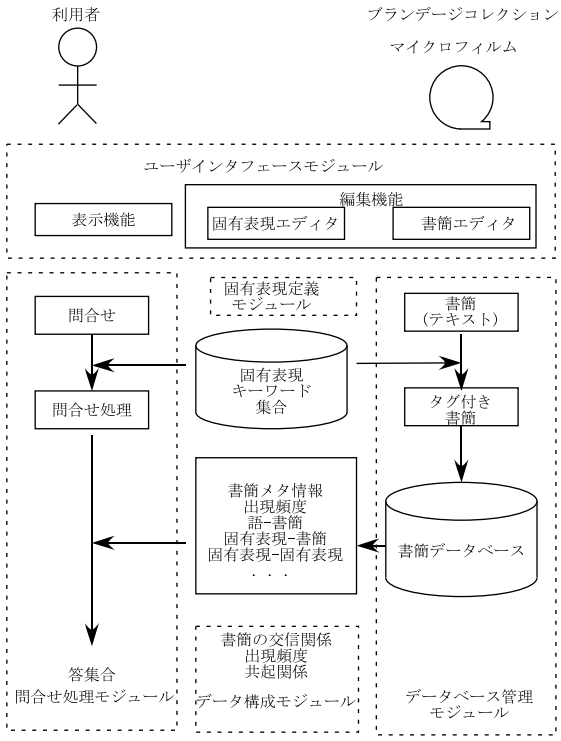


図 1 LMBC の概念的構造
Fig. 1 A conceptual overview of LMBC.

[*3] 固有表現とキーワードやキーフレーズは同じ方法によって定義されている．本稿では，混乱が生じないならば固有表現はキーワードを含むとする．

られている．個々の書簡は 1 ページ（用紙 1 枚）ずつ，1 コマのフィルムに写されており，ボックスの区切にはボックスのイメージが利用されている．ただし，1 つの書簡としての区切りは示されていない．LMBC では，個々の書簡はマニュアル操作によってテキスト形式で保存され，構成要素がタグ付けされる．この手続きにより，書簡はタグを用いて XML 形式で記述される．タグには 2 種類がある．1 つは書簡の構造記述のためのタグであり，他方は固有表現を示すタグである．書簡の XML による記述の詳細は，5.1 節に述べる．

現在，データベースは，1940 年に開催予定であったオリンピックに関する書簡，女子スポーツに関する書簡，および Sigfrid Edström（ジークフリード・エドストレーム）（第 4 代 IOC 会長）と交換した書簡などよりなる．主な書簡の内容は，意見交換，会議への招待，受取人の紹介や推薦，旅程の通知などであり，書簡は簡潔に記述されていることが多い．

データ構成モジュールは，問合せを処理するために必要な内部データを構成する．書簡に出現する語彙の集まり，固有表現の共起関係などを，内部データとして構成する．

ユーザインタフェースモジュールは，大きく 2 つのモジュールからなる．1 つは編集機能を実現する編集モジュールであり，他方は種々の表示機能を実現する表示モジュールである．編集モジュールは，書簡を定義・修正するための書簡エディタ，および固有表現を定義・修正するための固有表現エディタよりなる．また表示モジュールは，問合せに対する答集合から，書簡交換，本文中の固有表現や語彙の出現頻度などを表示する．なお互いに関連する固有表現を得るモジュールは，問合せ処理モジュールの一部である．

固有表現やキーワードは，固有表現エディタを用いて定義される．固有表現の記述は，(1) 固有表現を表し，その同義語の代表となる代表語，(2) 固有表現の種類，(3) 固有表現の同義語のリストからなる．固有表現の種類には，一般に設定されている人名，地名，日付，領域固有の種類として設定されている会議名，組織名，大会名，スポーツ名である．また，キーワード，キーフレーズおよび “固有単語” と呼ぶ固有名詞を種類として設けている．固有単語には，分類することが難しい固有名詞が定義される．たとえば，新聞名，船名などである．

固有表現の XML 表現を直接操作することは難しいので，固有表現エディタは固有表現の階層構造にしたがって操作するように設計した．すなわち，固有表現の種類–代表語–同義語という 3 つのレベルで，個々のレベルで定義内容を操作する．たとえば，新しい語彙または文字列を固有表現として定義するとき，まず種類を選択する．代表語が定義されていなければ代表語を定義する．代表語の下で同義語となる語彙を同義語リストに設定する．このとき，書簡に記載された表記が分析対象となることがあるので，表記どおりに同義語リストに設定する．たとえば，文字列の大文字や小文字，英語と米語の違いなどを区別する．また，新しく固有表現や同義語を定義したときには，重複をチェックするようにした．さらに，現在固有表現の種類は固定しているが，書簡が増加すれば設定している種類の不足による種類の増加や，種類の詳細化が考えられる．

問合せモジュールは利用者の記述する問合せに応えるためのモジュールである．書簡の構成要素である差出人，受取人や日付を用いた検索，語彙の集まりによって本文検索を可能とすることが適切であると考えている．また，差出人や受取人は代表語による検索や表示が適切である．たとえば書簡交換を示すとき，差出人と受取人が代表語で表されれば理解が容易である．しかし同義語が設定さている差出人・受取人を代表語で表すと，書簡に表記された氏名と代表語とが異なることがある．このため人名の代表語を，一般に通行していると考えられる呼称としての代表語とする．さらに，語彙の集まりによる問合せの記述では，検索条件となる語彙を与えたときにその代表語や同義語による検索を容易にする．これらは，固有表現定義モジュールを利用して処理可能である．利用者は，同義語を反映する検索であるのか，反映しない検索であるのかを指定する．同義語検索が指定されたとき，同義語検索は次のように実行される．問合せに与えられた語彙の代表語を求める．代表語によって本文はタグ付けされているので（5.2 節参照），代表語によるパタンマッチングによって，同義語を含むすべての書簡が検索できる．

LMBC は Java を用いて実装されている．本システムを実装するためにいくつかのパッケージを用いた．それらは，文を語彙に分割するための形態素解析ステム TreeTagger [31]，固有表現共起ネットワークを処理するための GraphStream [32] である．表示モジュールにおいては，3 次元表示のためにグラフィックライブラリ Java3D，ネットワーク表示のために GraphViz [33]，および出現頻度のグラフを作成するために JFreeChart [34] を用いた．

5. 書簡の表現

5.1 書簡のデータ構造

個々の書簡は，4 種類の形式で書簡データベースに格納されている．それらは，(1) マイクロフィルムから得られた PDF 形式の画像ファイル，(2) 画像ファイルから書簡部を読み取ったテキストファイル，(3) 書簡の構成要素や本文の段落がタグ付けされた，段落タグ付き文，および (4) 段落タグ付き文から自動的に固有表現がタグ付けされた，固有表現タグ付き文である．

書簡を定義するために設けたタグを**表 1** に，XML 形式による書簡の一部を**図 2** にそれぞれ示す．書簡の集まりを定義するルートノードは `letters` である．個々の書簡

表 1 書簡記述のために設けたタグ一覧
Table 1 Tags for describing letters.

タグ	要素
letters	書簡の集まり
letter	1 つの書簡．書簡の識別子を記入する
	属性 id が定義されている．
collection-box	書簡が保存されているコレク
	ションボックスの番号
sender	差出人，代表語を記入する
	id が定義されている．
name	書簡に表記されている差出人名
organization	差出人の所属
address	差出人の住所
receiver	受取人．代表語を記入する
	id が定義されている．
name	書簡に表記されている受取人名
organization	受取人の所属
address	受取人の住所
date	差出日の日付．日付は標準化されて，
	yy-mm-dd の形式に変換される．
day	日
month	月
year	年
subject	主題が記されていれば記入
text	段落タグ付き文．書簡の本文
p	段落タグ付き文を構成する段落
body	固有表現タグ付け文
	段落タグ付き文の固有表現をタグ付け
p	固有表現タグ付け文を構成する段落
copy-receiver	同報された受取人リスト
carbon-copy	同報された同報者リスト
blind-carbon-copy	受取人に同報を伏せた同報者リスト
original	書簡のオリジナルテキストまたは画像
text	書簡本文のテキスト化の結果
pdf	書簡の画像である PDF ファイル

は，letters の子ノードとして要素 letter に記述される．letter の属性 id には，システムが管理する書簡の識別子が設定される．個々の書簡の構成要素は，差出人，受取人，日付，主題，段落タグ付け文，固有表現タグ付け文，コピーの受取人および元の書簡ファイル名を記述するオリジナルよりなる．それぞれの要素のタグは sender，receiver，date，subject，text，body，copy-receiver および original である．要素 sender と receiver は，name，organization および address よりなり，それぞれ書簡に表記された名前，所属機関および住所（所在地）である．差出人と受取人は表記とは別に，固有表現の代表語を用いて参照する必要がある．表記された名前 name から，人名の代表語が調べられる．代表語は sender と receiver に設けられている属性 id の属性値に記入される．date は日，月および年を記入するための要素 day，month および year が設定されている．記された日付から自動的に標準

```
-<letters>
...
-<letter>
-<letter id="L-66">
...
-<sender id="Mr. Brundage">
<name>Avery Brundage</name>
...
</sender>
-<receiver id="Mr. Edward W. Mumford">
<name>Mr. Edward  W. Mumford</name>
...
<date year="1938" standard="1938-04-19" month="April"
day="19"/>
-<text>
<p> It was a great pleasure to learn from your recent letter that the
University of Pennsylvania will confer an honorary degree on Mr. J.
Sigfrid Edstrom when he arrives in the United States this summer as
the head of the Swedish Delegation which comes to the United States
to assist in the celebration of the 300th Anniversary of the Swedish
Settlement of Delaware. </p>
...
<p> Sincerely yours, </p>                          段落タグ付文
</text>
-<body>
-<p>                                              固有表現タグ付文
It was a great pleasure to learn from your recent letter that
<OrganizationName> the University of Pennsylvania,the University
of Pennsylvania </OrganizationName>
will confer an honorary degree on
<PersonalName> Sigfrid Edstrom,Mr. J. Sigfrid Edstrom
</PersonalName>
when he arrives in the
<LocationName> United States,United States </LocationName>
this summer as the head of
<OrganizationName> the Swedish Delegation,the Swedish Delegation
</OrganizationName>
which comes to the
<LocationName> United States,United States </LocationName>
to assist in the celebration of the 300th Anniversary of
<OrganizationName> the Swedish Settlement of Delaware,the
Swedish Settlement of Delaware </OrganizationName>
.
</p>                                             下線部は固有表現
...
<p> Sincerely yours, </p>
</body>
-<original>
<pdf>L-66.pdf</pdf>
...
</original>
</letter>
...
</letters>
```

図 2 XML 形式による書簡記述の一部
Fig. 2 Description of a letter represented in XML form.

形の日付に変換される．標準形の日付は，yyyy-mm-dd であり，西暦年 – 月 – 日を表す．copy-receiver は，子ノードとしてカーボンコピーやブラインドカーボンコピーの受取人を記入する carbon-copy と blind-carbon-copy が設けられており，受取人となる人物名のリストが記入される．original は，PDF 画像ファイル名を記入する pdf，テキストファイル名を記入する text よりなる．このとき，書簡が複数のページからなれば，1 つのファイルにまとめられる．

段落タグ付き文から固有表現タグ付き文を得るために，段落付き文に出現する固有表現が調べられて，本文がタグ付けされる．図 2 においてタグ <text> と <body> に囲ま

れた部分は，それぞれ段落付きタグ付文，および固有表現タグ付き文である．また，固有表現タグ付き文において，タグ付けされた固有表現を下線で示す．ただし，下線は説明のために付した．なお，本文を構成する段落はタグ <p> により示される．

1つの書簡を定義するためには，少なくとも受取人，発信人，日付，段落タグ付文およびPDFファイル名を記入する．6章に述べる書簡の受取人・差出人間の書簡の交換を表すネットワークは，受取人，発信人および日付より構成される．また，7章に示す固有表現の共起関係は固有表現タグ付き文から生成される．

その他のタグは，書簡に記入された項目を定義するためや，分析機能の拡張に備えるために設けたタグである．たとえば，組織や所属を代表する意見をまとめる必要があれば，組織を代表する発信人，受取人を記述する sender および receiver の address や organization に記される情報が必要である．また，コピーの受取人リストは情報の共有を示すために必要である．

5.2 固有表現の処理

固有表現である人名，組織名，会議名，大会名などを書簡に記述するとき，正式名称やその略称が用いられる．また，差出人と受取人が共通に認識しているならば，固有表現の略称のみが記述されるときもある．たとえば，人名 Mr. Brundage は Avery Brundage，A. Brundage，など，また，組織名 The International Olympic Committee は IOC，I.O.C.，などと記される．**表 2** に固有表現の例を示す．

一方，書簡には主としてスポーツや競技に関する話題が記されており，オリンピックに関連する固有の語彙や，専門的な意味や固有の意味のある語彙が出現する．たとえば，Olympic charter，Olympic village，regulation，Woman，など．この種の語彙は，書簡の理解に欠かせない語彙であり，キーワードやキーフレーズである．

固有表現は，以下のようにタグ付けされる．

(1) タグは固有表現の種類を表している．たとえば，人名の PersonalName，キーワードの Keyword，など．
(2) 固有表現の種類を表すタグには，属性 id が定義されている．id の属性値は，固有表現の代表語である．
(3) 代表語の同義語はリスト形式で表される（同義語リスト）．同義語を定義するタグの形式は，下のとおりである．

```
<固有表現の種類>
    <EntityName id = 代表語> syn1, syn2, ···
    </EntityName>
    ···
</固有表現の種類>
```

代表語の同義語は，タグ tag のコンテンツ部に文字列として記入され，任意個の同義語が記入される．

表 2 固有表現の例
Table 2 Some examples of named entities.

代表語	同義語
人名	
Dr. Ohno	Seishichi Ohno, Dr. S. Ohno,
Mr. Brundage	Avery Brundage, A. Brundage, Mr. Avery Brundage
Dr. M. Nagai	Dr. Nagai, Matsuzo Nagai,
会議名	
Cairo meeting	The Olympic Congress at Cairo, the Cairo Session
組織名	
the Japanese Olympic Committee	the Japanese Olympic Committee, J.O.C., JOC Japan Olympic Committee
5th Olympic Winter Games Committee	Vth Olympic Winter Games Committee Sapporo, 5th Olympic Winter Games Committee
the International Olympic Committee	the I.O.C., IOC International Olympic Committee
地名	
United States	United States, U.S., America, U.S.A, USA

固有表現は代表語とその同義語リストによって表されている．まず，すべての代表語と同義語リストの要素が集められ，文字列の長さの順に並べられる．長さの長い文字列から順に書簡本文に含まれているかどうかが調べられる．もし調べている文字列が本文に含まれていれば，その文字列は固有表現としてタグ付けされる．

本文に出現する固有表現であるテキストは，上記の(1)〜(3)に示す定義を用いて下のように書き換えられる．

<tag> 代表語，出現語 </tag>

tag は固有表現の種類を表す．出現語は段落タグ付け文に出現する語彙である．たとえば，出現語 U.S.A. は，固有表現タグ付け文ではタグを用いて <LocationName> United States，U.S.A.</LocationName> と書き換えられる．タグ LocationName に囲まれたコンテンツ部は，代表語が United States であり，種類が地名 LocationName であることを表す．

6. 書簡の表示

文書処理のための基本機能として，LMBC は (1) および (2) に示す機能を提供している．(1) 書簡と書簡に出現する語彙を表すネットワークを表示する．共通の話題について述べられた書簡が示される．(2) 共起する語彙の共起ネットワークを示す．ある話題に関して述べられるとき，共起する語彙の集まりを示す．

一方，書簡の表示では，以下に示す書簡や語彙間の関連

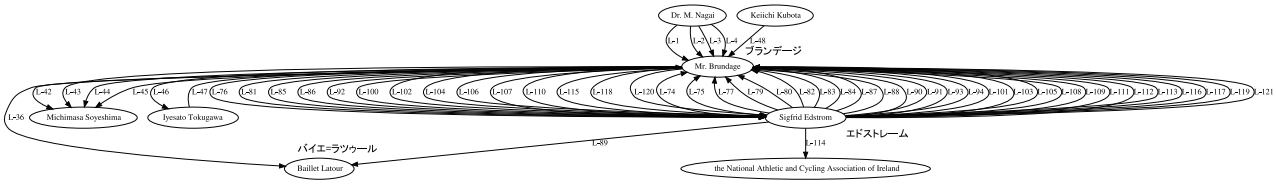


図 3 書簡交換ネットワークの例
Fig. 3 An example of a letter exchange network.

を示す必要がある．

- 受取人，差出人および期間などを指定して，差出人と受取人との書簡交換を示すネットワークを示す．
- 書簡の集まりをウィンドウに表示するとき，時間軸に沿い書簡の日付順に書簡を表示する．このとき，ウィンドウに表示できる期間や，時間の単位を指定する．
- 個々の書簡に出現する語彙の出現頻度を示す．このために 1 つの書簡を棒グラフによって記述する．
- 時間的に連続する書簡に出現する語彙を表示して，書簡が記述する内容の理解に供する．

本節では，上に述べた 4 つの要求に応じて開発された 2 つの機能について述べる．それらは，書簡の交換を示すネットワークを表示する機能と，個々の書簡に出現する語彙の出現頻度を示す棒グラフを表示する機能とである．

6.1 書簡交換の表現

分析では，書簡の内容とともに，差出人と受取人との書簡交換を調べたり，差出日順に並べたりする．書簡交換をネットワークによって表示することにより，書簡交換の直感的な理解が容易となる．この要求に応えるために，本システムは，以下に示す 2 種類のネットワーク表現を提供している．

(1) 書簡交換ネットワーク．ネットワークのノードは受取人または差出人である．リンクは差出人から受取人への有向リンクであり，リンクのラベルは書簡の識別子である．**図 3** は，1935 年～1938 年に交わされた 45 通の書簡交換を表す書簡交換ネットワークである．たとえばブランデージは，Le comte Henri de Baillet-Latour（アンリ・ド・バイエ=ラトゥール伯爵）（第 3 代 IOC 会長），およびエドストレームと書簡を交換したことが示されている．ただし，図中のカタカナによる人名表記は注釈のために記入した．

(2) 時間順（日付順）書簡交換ネットワーク．時間順書簡交換ネットワークは，書簡交換を 3 次元空間に構成する．時間順書簡交換ネットワークを構成するためのデータ構造を**図 4** に示す．まず，書簡は時間順に等間隔に整列される．日付の新しい書簡が上に位置する．次に，表示対象となる書簡の差出人と受取人となっている人物の集まりを求める．人物の表示位置を，上辺の円周上に等間隔に配置する．さらに，個々の書簡の差出人から受取人へのリンクが円柱の円周上に示される．差出人から受取人への距離が短い方の円周上の弧に沿って，リンクが設定される．リンクのラベルは，識別子と差出日の日付である．時間順書簡交換ネットワークの例を**図 5** に示す．利用者は，3 次元空間上の視点や位置を変更することにより，時間軸を移動したり，角度を変えたりして書簡交換を確認できる．

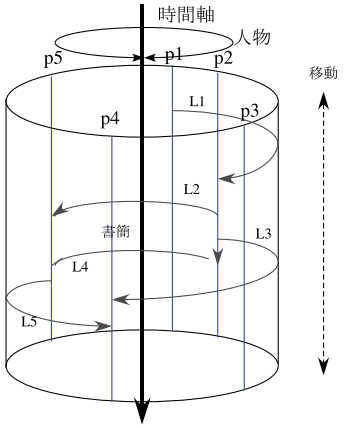


図 4 時間順書簡交換ネットワークのデータ構造
Fig. 4 Data structure of time-ordered letter exchange network.

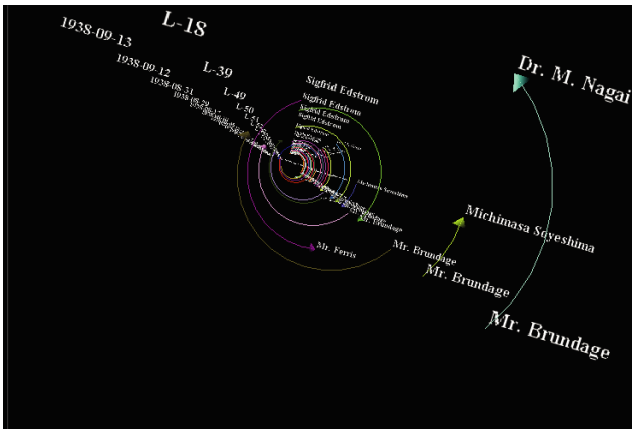
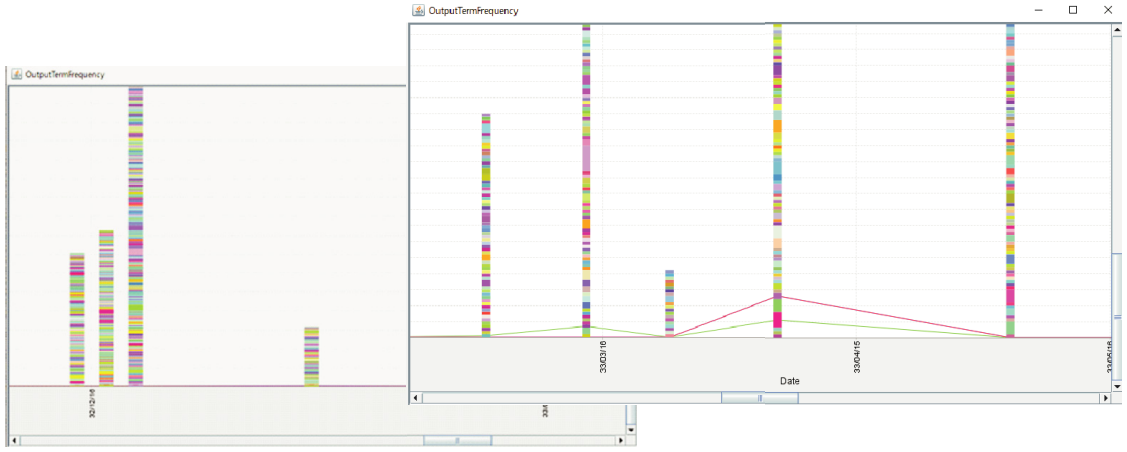


図 5 時間順書簡交換ネットワークの一部
Fig. 5 A part of time-ordered letter exchange network.

6.2 書簡の棒グラフによる表現

語彙の出現頻度を棒グラフによって表示する方法は，既存の文書分析システムにおいて実装されている [16], [30]．LMBC では書簡に相当する棒グラフは，時間順に整列され



(a) (b)

図 6 書簡における語彙の出現頻度を表す出現頻度グラフの一部
Fig. 6 A part of a word frequency graph for letters.

る．このとき，書簡の送受信間隔は一定ではないので，交換の頻度を示す必要がある．

上記の要望に応じるために，書簡の表示のために作成するウィンドウに表示できる期間と，時間軸の単位を指定できるようにした．たとえば，書簡の集まりを表示するとき，ウィンドウに 3 年間分の書簡を表示して，時間軸の目盛りを月単位とするというように指定する．なお，表示対象である書簡の集まりが 3 年間に収まっていなければ，次の 3 年間の書簡を表示するウィンドウが作成される．

図 6 に語彙の出現頻度を表す出現頻度グラフを示す．横軸は時間軸であり，縦軸は語彙の出現頻度である．書簡に出現する個々の語彙は，異なる色で示される．また異なる書簡に出現する同じ語彙を，要求に応じて線により結ぶ．図 6 (a)，(b) は，表示画面における表示時間間隔と時間単位を変更した例である．また図 6 (b) における単語を結ぶ直線は，指定した語彙が出現している書簡を示している．ウィンドウに示されている他の書簡に同じ語彙が出現すれば，単語は直線によって結ばれる．

7. 共起関係に基づく固有表現間の関連と実験

LMBC 開発の目的の 1 つは，書簡に出現する固有表現間の関連を提示して支援することである．ある 2 つの固有表現が与えられたとき，それらを互いに関連付ける固有表現を求める．固有表現の共起関係は，共起する固有表現は互いに関連していることが多いであろうという関係である．2 つの固有表現間に存在する固有表現を，関連する固有表現と見なす．得られた固有表現の集まりを関連固有表現と呼ぶ．5.2 節に示したように固有表現は，キーワードやキーフレーズを含んでおり，それらは固有表現共起ネットワークの構成要素である．

7.1 2 つの固有表現間の関連を求める手順

固有表現共起ネットワークのノードは固有表現であり，媒介中心性とクラスタ係数に基づいて関連する固有表現を求める．媒介中心性は，あるノードが他の 2 つのノードの最短経路上のノードである割合である．媒介中心性が高いノードは，任意の 2 つのノードが表す固有表現を結び付けるための固有表現であると見なされる．また，クラスタ係数は，あるノードと直接関連付けられた他のノードが互いに関連づけられている割合である．クラスタ係数の高い固有表現は，隣接する固有表現と関連が深い固有表現である．

固有表現共起ネットワークのノードとリンクは，それぞれ代表語と，固有表現の共起関係である．1 つの書簡本文に 2 つの固有表現が出現するとき，これらの固有表現にリンクが設けられる．隣接する 2 つのノード間の距離は 1 である．本実験では，共起関係のみに注目しており，ネットワークの構造に基づく関連付けメカニズムを実装している．

実験では，約 170 通の書簡から約 330 個の固有表現やキーワードの代表語が得られた．得られた固有表現の集まりから固有表現共起ネットワークを構成した．個々の固有表現の媒介中心性とクラスタ係数との分布を，**図 7** に示す．媒介中心性とクラスタ係数には，一方の値が高いと，他方の値は低いという関係がある．なお，媒介中心性とクラスタ係数の平均は，それぞれ 0.71×10^{-2} および 0.80 である．媒介中心性が最も高い固有表現ノードは，`United States` であった．また，図 7 における (a)，(i)，実線および破線については，7.2 節に示す．

関連固有表現を求める手順は，次のとおりである．まず，媒介中心性およびクラスタ計数に関する 2 つの閾値，それぞれ th_b および th_c を設ける．また，固有表現 n の媒介中心性とクラスタ係数を，それぞれ $b(n)$ および $c(n)$ とする．

(1) 与えられた 2 つの固有表現（代表語）間を結ぶ最短経路集合を求める．ただし，与えられた 2 つの固有表現

が直接結ばれているとき，または最短経路がないときには終了する．

(2) 個々の最短経路に関して，下の手順を実行する．

(a) 最短経路上にあるすべての固有表現（与えたノードは除く）が，媒介中心性またはクラスタ係数のどちらかの閾値を満たしていれば，すなわち $b(n) \geq th_b$ または $c(n) \geq th_c$ ならば，n を関連固有表現の候補に加える．

(b) 最短経路上に閾値を満たしていない固有表現が1つでもあれば，その最短経路上の固有表現は関連固有表現には加えない．ただし，他の経路から得られた候補となっている固有表現は，候補からは除かない．

(c) 最短経路がなくなるまで，(a)，(b) を繰り返す．

(3) (1) および (2) によって得られた関連固有表現を，2つの固有表現の関連固有表現とする．

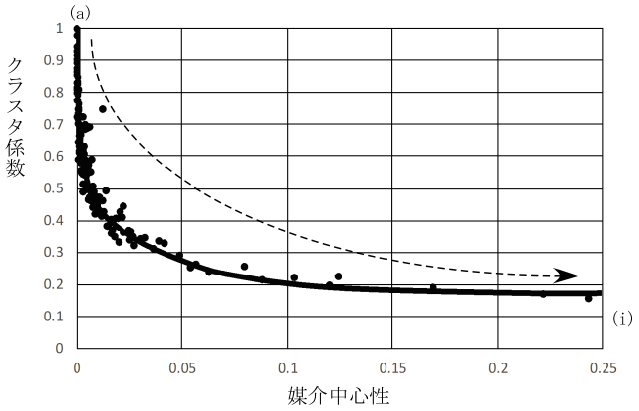


図 7 固有表現共起ネットワークにおける固有表現の媒介中心性とクラスタ係数の分布

Fig. 7 Distribution of betweenness and clustering co-efficient of named entities in a collection of letters.

なお，固有表現共起ネットワークの処理のためにグラフ処理パッケージ GraphStream [32] を用いている．固有表現タグ付け文から固有表現の共起関係を求めて，GraphStream のためのネットワークを構成した．固有表現の媒介中心性，クラスタ係数や，固有表現間の最短経路は GraphStream を用いて求めた．なお，最短経路は，GraphStream が提供するダイクストラのアルゴリズムを用いて求めた．

図 8 に2つの固有表現 `Olympic Movement` と `Japanese` との関連固有表現を示す．2重楕円形は問合せとなった固有表現を，楕円形は関連固有表現の要素を示す．なお，この図に示した例では，媒介中心性およびクラスタ係数に関する閾値は，それぞれ $th_b = 0.1$ と $th_c = 0.2$ である．

7.2 実験

対象とした約170通の書簡から得られた約330個の固有表現をノードとする固有表現共起ネットワークに対して，2つの固有表現の対である15組の問合せを与えて関連固有表現を得た．得られた関連固有表現の集まりの再現率と精度とを求めた．15組の問合せを**表 3** に示す．問合せは固有表現の代表語である固有名詞やキーワードである．問合せで用いた固有表現の延べ数は，問合せが15組であるので30種であり，異なり数は20種である．なお，想定した答となる固有表現は，対象とした書簡の集まりから適切であると考えられた固有表現である．

また，9組の閾値を設定した．実験では閾値を変更して，問合せ実験を行った．閾値の組を**表 4** に示す．閾値は図7に示した媒介中心性とクラスタ係数の分布にしたがって閾値の組を設定した．媒介中心性とクラスタ係数を変化したとき，得られる関連固有表現の精度と再現率がどのように

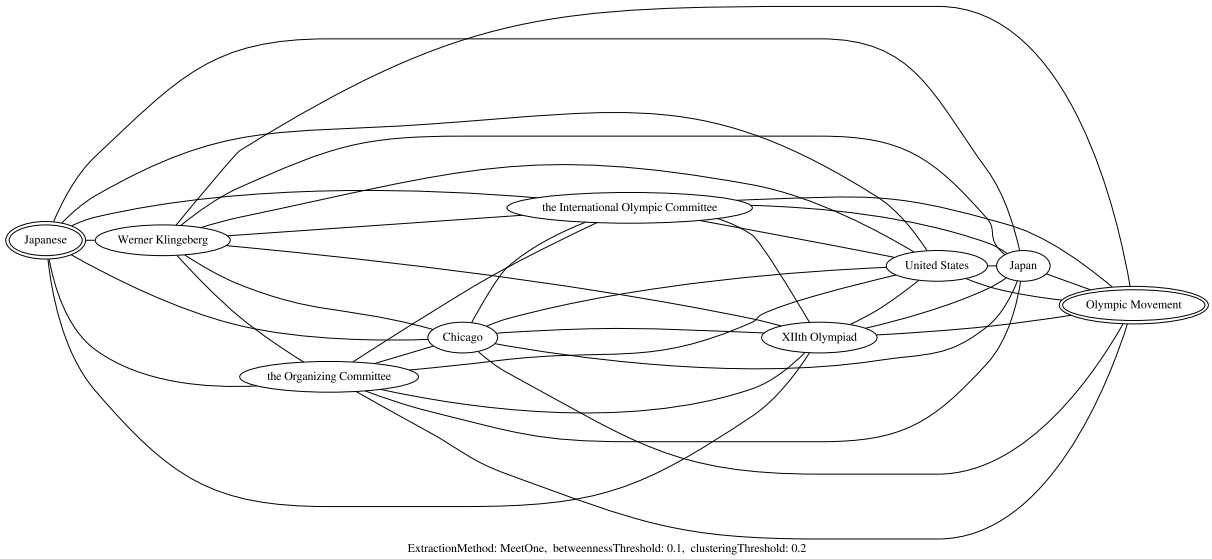


図 8 固有表現 `Olympic Movement` と `Japanses` とを関連付ける関連固有表現の例

Fig. 8 Examples of named entities for connecting two named entities `Olympic Movement` and `Japanese`.

表 3 実験で用いた 15 組の固有表現とキーワード
Table 3 15 cases of named entities and/or keywords for experiments.

	問合せ対	正解と想定した固有表現とキーワード
1	Germany, Tokyo Olympiad	Tokyo, War, Berlin, Chicago, United States, Europe
2	Woman, European Championships	Berlin, Mr. Brundage, Stockholm
3	Baron M. Inada, Paris	Cairo, Chicago, United States, the International Olympic Committee
4	Woman, European	Amateur Athletic Union, American, Berlin, Gymnastics, Mrs. Edstrom, Olympic Games, Olympic, Stockholm, Track and Field, swimming, ski, the International Olympic Committee
5	Warsaw Session, ski	Olympic, Tokyo, XIIth Olympiad, the Cairo Session
6	amateurism, Japanese	Mrs. Brundage, Olympic Games, United States, United States Olympic Association, the International Olympic Committee
7	Japanese, the Olympic Games in Berlin	American, Council, United States, the International Olympic Committee
8	Sigfrid Edstrom, Baron da Coubertin	Brindisi, Cairo, Olympic Games, United States
9	Sigfrid Edstrom,	London Congress, Warsaw, the Cairo Session, the I.A.A.F. Congress, the International Olympic Committee, Coubertin heart Olympia
10	Tokyo, Brussels	Berlin, Council, General Sherrill, Olympic Games, United States Olympic Association, the Executive Committee, the International Olympic Committee, Oslo
11	Amateur Athletic, Coubertin heart Olympia	Amateur Athletic Union, Baillet Latour, Berlin, Chicago, Council, Greece, International Amateur Athletic Federation, the Cairo Session, the I.A.A.F. Congress, the International Olympic Committee, London Congress
12	Baron da Coubertin, Berlin	Amateur Athletic Union, Helsinki, Olympic, Olympic Games, XIIth Olympiad, the American Olympic Association, the Organising Committee
13	Baron da Coubertin, Woman	Amateur Athletic Union, Olympic, Olympic Games, United States, XIIth Olympiad, the Organising Committee
14	Olympic Movement, Amateur Athletic	Amateur Athletic Union, Berlin, Council, International Amateur Athletic Federation, XIIth Olympiad, the International Olympic Committee, the Organising Committee
15	Olympic Movement, Japanese	Japan, United States, XIIth Olympiad, the International Olympic Committee, the Organising Committee

表 4 媒介中心性とクラスタ係数に関する閾値の記号と閾値対
Table 4 Some thresholds of betweenness and clustering-coefficient.

記号	媒介中心性	クラスタ係数
(a)	0.00	1.00
(b)	0.15×10^{-2}	0.65
(c)	0.35×10^{-2}	0.60
(d)	0.06×10^{-2}	0.48
(e)	0.13×10^{-2}	0.41
(f)	0.25×10^{-1}	0.35
(g)	0.50×10^{-1}	0.25
(h)	0.10	0.20
(i)	0.25	0.10

変化するのかを調べるために，図 7 における実線で表される曲線上の値を閾値の対として与えた．図 7 に示すように，閾値 (b) から (h) は (a) から (i) の間に破線の方向に沿って順に位置している．7.1 節に示した手順に示したように，閾値 (a) では媒介中心の低い固有表現は検索対象ではない，また閾値 (i) ではクラスタ係数の低い固有表現は検索対象ではない．(a)〜(h) の間を媒介中心性の値が 1/2 となるように設定し，曲線上に対応するクラスタ計数の値を求めて，媒介中心性とクラスタ係数の対を閾値とした．

さらに実験では，比較のために 2 種類の閾値の比較方法を用意した．7.1 節に示した手順を MO と呼ぶ．MO によって得られた関連固有表現を比較するために，閾値の利用方法を変更して，次の (1) と (2) の方法を用意した．(1) 媒介中心性は閾値 th_b の値より高いが，クラスタ係数の値は th_c より低い（$b(n) \geq th_b$かつ $c(n) \leq th_c$）という方法，(2) 媒介変数は閾値より低いが，クラスタ係数は高い（$b(n) \leq th_b$かつ $c(n) \geq th_c$）という方法である．(1) は媒介変数を優先するので BP，(2) はクラスタ係数を優先するので CP とそれぞれ呼ぶ．

15 組の問合せを与えて，MO，BP および CP により得られる，関連固有表現の再現率と精度との分布を**図 9** に示す．MO，BP および CP で得られた再現率と精度の結果を，それぞれ記号●，■および × で表す．また，これらの

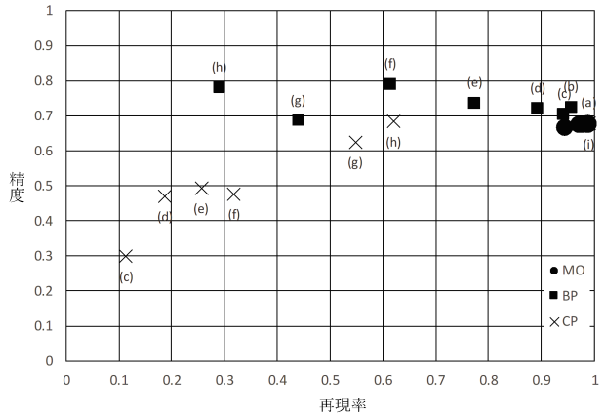


図 9 関連付けの方式を変更したときの再現率と精度の分布
Fig. 9 Distribution of recall and precision when relating method of named entities is changed.

記号の側に示した (a)〜(i) は，表 4 に示した閾値に対応する．ただし，MO のための記号●は重複が多いので，閾値の記号は示していない．

MO では，2 つの閾値のうち，1 つの閾値を満たすと関連固有表現としているので，閾値を変更しているにもかかわらず，調べられる固有表現の数は多い．媒介中心性とクラスタ係数の変化において，2 つの閾値を満たさない固有表現のみが調べられない．MO の再現率と精度はともに，変化が少なかった．15 組の問合せ例に対する再現率および精度は，それぞれ約 0.98 および約 0.67 であった．

また BP では，固有表現の媒介中心性の値が，閾値を満たす固有表現のみが調べられる．閾値を変更すると，精度の変化は少ないが，閾値の変化に応じて再現率の値は変化する．媒介中心性の値が高くなると，調べられる固有表現が少なくなる．このために，精度の変化は少ないが再現率は低くなる．すなわち，th_b を高くすると，再現率が低下する．媒介中心性は，再現率との関連が深いと考えられる．

一方 CP では，クラスタ係数の値が，閾値を満たす固有表現のみが調べられる．CP による再現率と精度は，BP の再現率と精度に比較して低い．クラスタ係数の閾値を高くすると，調べられる固有表現は少なくなる．th_b を高くして，th_c を低くするに従い，精度と再現率はともに向上する．調べられる固有表現が増えることにより，再現率は向上する．しかし，同時に精度も向上している．これは閾値が (a) から (i) に変更するに従って，調べる固有表現が増えるので，MO の結果に近づくためである．

なお MO による関連付け結果の再現率は高い．また，図 9 に示すように，閾値が変化しているにもかかわらず精度再現率の変化が少ない．閾値 (a) から (i) への変化にしたがってクラスタ係数の低い固有表現が関連固有表現として検索されなくなったとしても，媒介中心性の値の大きい固有表現が関連固有表現の候補となるので，再現率と精度が大きく変化することがなかったのであろう．

8. 考察

本システム開発の目的は，関連のある書簡を表示することと，固有表現を関連付けて書簡の理解を支援することであった．6 章に示した書簡交換ネットワーク，時間順書簡交換ネットワークや，書簡に出現する用語の頻度を表す棒グラフによって書簡間の関連が示される．個々の書簡の交換や，時間軸に沿って同じ語彙を含む書簡を表示することで，直感的な理解の容易性を高め，書簡を探すための負担の軽減が図られると考えられる．

また，固有表現間の関連付けでは固有表現共起ネットワークを構成して，関連固有表現の再現率と精度を求めることによって，関連付けの方法と結果とを評価した．再現率が我々の実験では 0.98 であり，2 つの固有表現を関連付ける候補となる固有表現は，我々が対象とした書簡集合から固有表現やキーワードの共起関係に基づき，提案アルゴリズムによって適切に得られている．一方，システム利用者の観点からは，関連固有表現を提示する可能性が高いので，再現率に関する結果は有用であると考えている．利用者は，得られた関連する固有表現が興味深ければそれを検討するであろう．したがって再現率が高いことは，新たな関連を調べたり確認したりする対象が示唆されることになるであろう．

固有表現は特定の概念や人物などの対象を表す．書簡では，差出人と受取人の間で対象の理解が共有されていることが多い．共起関係に基づいて 2 つの対象間の関連を得ることは，直接的に表されていない対象間の関係を得る方法の 1 つであると考えている．

書簡テキストの集まりを定義するという観点から，今後の課題として利用者負担の軽減があげられる．書簡データベース構築時に，書簡のテキスト作成に関する負担が大きい．画像ファイルから書簡をテキスト化するとき，視覚的な特徴（書簡表現の領域）に基づき必要なテキストを自動抽出できれば，負担は軽減される．このためには，たとえば書簡の受取人や差出人の記述位置や名前，住所，段落などを抽出する必要がある．また，5.2 節に示したように，固有表現の設定は開発者または利用者の負担である．この負担を軽減するために固有表現抽出ツールの利用が考えられる．したがって今後，スポーツ・オリンピック分野に対する辞書の整備が必要である．さらに一部の画像が可読ではなく，スキャナによる読み取りができない場合がある．

9. おわりに

本稿では，史料である書簡の解読支援を目的とした LMBC のアーキテクチャおよび基本的な機能について述べた．それらは，書簡の特徴に基づく表示機能と，固有表現共起ネットワークに基づく固有表現の関連付け機能である．これらの機能を備えたプロトタイプシステムを構築して，LMBC

は開発の第1段階は，終了したと考えている．

現在本システムの実用性を高めるために，書簡の集まりを既存のデータベース管理システムに格納することを試みている．また今後，対象となる書簡や固有表現を増やして，スポーツ・オリンピック史研究の観点から分析結果やシステムを評価することを予定している．

謝辞 本研究の一部はJSPS科学研究補助金16H01867，および中京大学戦略的研究「スポーツ・デジタルアーカイブズ共同研究」の助成を受けた．有益なご意見をいただいた査読者の方々に深謝いたします．

参考文献

[1] 來田享子：スポーツ・デジタルアーカイブとスポーツ教育の未来，デジタルアーカイブ学会誌，Vol.4, No.3, pp.260–264 (2020).
[2] Brichford, M.J., fur Sportwissenschaft, B. and Brundage, A. (Eds.): *Avery Brundage Collection 1908-1975*, Verlag Karl Hofmann Schorndorf (1977).
[3] University of Illinois Archives: Avery Brundage Collection, 1908-82, available from ⟨https://archon.library.illinois.edu/index.php?p=collections/controlcard&id=4719&q=Brundage⟩ (accessed 2020-11-25).
[4] The International Centre for Olympic Studies at Western University: Avery Brundage Collection, available from ⟨https://www.uwo.ca/olympic/collections/brundage.html⟩ (accessed 2021-04-23).
[5] Ito, H., Kasugai, K., Kanamori, S., Kanekasu, S., Tashiro, S., Taki, T., Hasegawa, J., Raita, K.: Structure of a Prototype System for Managing Letters, *Procedia Computer Science*, pp.11–18 (2014).
[6] Ito, H., Kasugai, K., Kanamori, S., Kanekasu, S., Tashiro, S., Taki, T., Hasegawa, J. and Raita, K.: Retrieval and Analysis Mechanisms in Letter Management System, *Proc. 11th Intl. Conf. Knowledge Management*, pp.104–113 (2015).
[7] Ito, H., Miyazato, K., Ishikawa, K., Taki, T., Hasegawa, J. and Raita, K.: Analyzing Elements of Letters in a Letter Management System, *4th Intl. Conf. Applied Computing and Information Technology*, pp.32–37 (2016).
[8] Palshikar, G.K.: Techniques for Named Entity Recognition: A Survey, *Collaboration and the Semantic Web: Social Networks, Knowledge Networks, and Knowledge Resources*, Bruggemann, S. and D'Amato, C. (Eds.), pp.191–217, IGI Global (2012).
[9] 岩倉友哉，関根 聡：情報抽出・固有表現抽出のための基礎知識，近代科学社 (2020).
[10] Witten, I., Bainbridge, D. and Nichols, D.: *How to Build a Digital Library*, 2nd ed., Morgan Kaufmann (2010).
[11] フランス国立図書館：Gallica, Bibliotheque nationale de France digital library, 入手先 ⟨https://gallica.bnf.fr/⟩ (参照 2021-04-22).
[12] 早稲田大学図書館特別資料室：古典籍総合データベース，入手先 ⟨https://www.wul.waseda.ac.jp/kotenseki/about.html⟩ (参照 2021-04-22).
[13] University of Cambridge: Cambridge Digital Library, available from ⟨https://cudl.lib.cam.ac.uk/⟩ (accessed 2021-04-22).
[14] University of Leeds, Library: Letters Database, available from ⟨https://library.leeds.ac.uk/special-collections/collection/709⟩ (accessed 2021-04-22).
[15] Whittaker, S., Bellotti, V. and Gwizdka, J.: Email in Personal Information Management, *Comm. ACM*, Vol.49, No.1, pp.68–73 (2006).
[16] Viégas, F., Golder, S. and Donath, J.: Visualizing Email Content: Portraying Relationships from Conversational Histories, *Proc. SIGCHI 2006 Conference on Human Factors in Computing Systems*, pp.979–988 (2006).
[17] Kang, H., Plaisant, C., Elsayed, T. and Oard, D.: Making Sense of Archived E-mail: Exploring the Enron Collection with NetLens, *Journal of the American Society for Information Science and Technology*, Vol.61, No.4, pp.723–744 (2010).
[18] Keim, D.A., Krstajic, M., Rohrdantz, C. and Schreck, T.: Real-Time Visual Analytics for Text Stream, *IEEE Computer*, Vol.46, pp.47–55 (2013).
[19] Tyler, J., Wilkinson, D. and Huberman, B.: Email as Spectroscopy: Automated Discovery of Community Structure within Organization, *The Information Society: An International Journal*, Vol.21, No.2, pp.143–153 (2005).
[20] Fortunato, S.: Community Detection in Graph, *Physics Reports*, Vol.486, pp.75–174 (2010).
[21] Li, X., Liu, B. and Yu, P.: Discovering Overlapping Community of Named Entities, *Proc. PAKDD*, pp.593–600 (2006).
[22] Faloutsos, C., McCurly, K.S. and Tomkins, A.: Fast Discovery of Connection Subgraphs, *Proc. KDD '04* (2004).
[23] Fang, L., Sarma, A.D., Yu, C. and Bohannon, P.: REX: Explaining Relationships between Entity Pairs, *Proc. VLDB Endoment*, Vol.5, No.3 (2012).
[24] Seufert, S., Berberich, K., Bedathur, S., Kondreddi, S.K., Ernst, P. and Weikum, G.: ESPRESSO: Explaining Relationships between Entity Sets, *CIKM '16*, pp.1311–1320 (2016).
[25] Zhang, K., Zhu, K.Q. and Hwang, S.: An Association Network for Computing Semantic Relatedness, *Proc. AAAI*, pp.593–599 (2015).
[26] Pirrò, G.: Building Relatedness Explanations from Knowledge Graphs, *Semantic Web Journal*, Vol.10, No.6, pp.963–990 (2019).
[27] Newman, M.: *Networks An Introduction*, Oxford University Press (2010).
[28] 坂本昭二，岡田至弘：古文書料紙の化学分析データベースの構築に向けて，情報処理学会研究報告，人文科学とコンピュータ研究会，Vol.2015-CH105, No.1, pp.1–6 (2015).
[29] 和所泰史，來田享子，木村吉次：戦後日本の国際スポーツ界復帰に関する永井松三の役割，スポーツ科学研究，Vol.35, pp.27–39 (2013).
[30] Graham, S., Milligan, I. and Weingart, S.: *Exploring Big Historical Data, The Historian's Macroscope*, Imperial College Press (2016).
[31] The Institute for Computational Linguistics of the University of Stuttgart: TreeTagger, available from ⟨https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/⟩ (accessed 2020-11-25).
[32] graphstream-project.org: GraphStream, available from ⟨https://graphstream-project.org/⟩ (accessed 2020-11-25).
[33] GraphViz.org: GraphViz, available from ⟨https://graphviz.org/⟩ (accessed 2020-11-25).
[34] JFree.org: JFreeChart, available from ⟨http://www.jfree.org/jfreechart/⟩ (accessed 2020-11-25).

伊藤 秀昭 （正会員）

1984 年東京電機大学大学院修士課程修了．（財）日本情報処理開発協会を経て，現在，中京大学工学部情報工学科教授．博士（工学）．人工知能システム，情報検索システム，デジタルライブラリ等の研究に従事．電子情報通信学会，IEEE，ACM 各会員．

來田 享子

中京大学大学院体育学研究科体育学専攻修了．現在，中京大学スポーツ科学部教授．中京大学スポーツミュージアム副館長．博士（体育学）．日本スポーツとジェンダー学会会長，日本体育・スポーツ・健康学会副会長，体育史学会副会長，国際オリンピック史家協会（ISOH）会員等．スポーツ史研究に従事．