伝統的モンゴル文字文書の時代横断型検索手法の提案

ハルタルフー ガルマーバザル
立命館大学 グローバル COE 研究補助員

前田 亮
立命館大学 情報理工学部

本論文では，800 年の歴史を持つ伝統的なモンゴル文字で書かれた文化遺産である古文書のディジタル図書館を構築することで，貴重な文献を活用するシステムについて述べる．伝統的モンゴル文字文書のディジタル図書館システムでは，モンゴル国内での有用性を考え，Unicode 化された伝統的モンゴル文字の文書を現代モンゴル語で検索することを可能にする時代横断型の検索手法を提案し，Greenstone ディジタル図書館システムに適用し，実装した．研究対象には，1604 年頃に書かれたとされているモンゴルの皇帝たちについて書かれた「アルタン・トブチ」という歴史書を用いた．利用者評価実験の結果，キリル文字入力を使用し検索を行う手法が大変よかったという点からシステムとしては高い評価が得られた．

An Approach to Cross-Age Information Retrieval of Traditional Mongolian Text

Garmaabazar Khaltarkhuu
Research Assistant
Global COE
Ritsumeikan University

Akira Maeda
College of Information
Science and Engineering,
Ritsumeikan University

This paper discusses possibilities to create a digital library on traditional Mongolian script. Also we introduce system architecture of a digital library that will store books and materials of historical importance written in traditional Mongolian which contain history of 800 years. Specifically, we propose a technique that will enable digital library system to allow users to search traditional Mongolian texts with keywords in modern Mongolian Cyrillic characters. To accomplish this goal, we use Greenstone digital library system and a traditional Mongolian digital library with Golden History chronological book of ancient Mongolian Kings and their history. Experiment result shows that this system obtained higher user evaluation since retrieval technique was used Cyrillic alphabet input.

1. まえがき

近年，インターネットの普及によりディジタル化された情報がだれにでも利用できるようになった．その中でインターネットと図書館を結びつけたディジタル図書館が注目されている．様々な文化的資料のディジタル化や保存に関する研究が行われている．しかし，古い文書は現代の文書に比べ，ディジタル化されているものが多くはない．

現在，モンゴル国立図書館の蔵書は約 400 万冊である．その中の一部には手書きの資料がある．モンゴル国立図書館のモンゴル研究・マニュスクリプト書庫では約 50,000 冊の非常に貴重な資料が保存されている．その中の約 21,100 冊の蔵書が 13 世紀から 17 世紀にかかわる文献である．しかし，ほとんどの資料が蔵書登録されておらず，登録されていてもタイトルが実際の本のタイトルと違ったり，本のタイトルが不明であり，保存や管理が不足しているなどの問題が残っている[1]．社会主義時代であった'90 年代までは，モンゴル史を研究する限られた人数の研究者しか利用していなかったが，最近では利用者が増えつつある．

一方，このような貴重な資料のモンゴル国立図書館での保存状態が悪いため，利用・公開することが不可能になり，直ちに解決するべき問題となっている．1951 年に建築された図書館の建物は室内の温度・湿度を調節できず，書庫内の湿度を上げるために，水がはられている「たらい」が数箇所に置かれているのが現状である．

一方，モンゴルでも書籍のデータベース作成，保存管理に関してコンピュータによる最新技術を取り入れる動きが始まった．モンゴル国政府の方針では，利用者レベル要求として，使いやすいこと，直ちに結果が出ること，そして現在モンゴルで一般的に使われているオペレーティングシステムにて動作し，さらにそのオペレーティングシステムの更新バージョンにて動作することを条件としている．

この論文では，モンゴル文字で書かれた文書のディジタル図書館の構成について述べる．このシステムでは，伝統的モンゴル文字の文書を現代モンゴル語で検索する手法を提案する．ま

た，Unicodeにおける基本文字から表示用文字に変換するコード変換手法を提案する．これらの手法を，Greenstone ディジタル図書館システムに実装した．実験の結果，基本変換条件に対応した検索語を用いた検索を，ほぼ問題なく行なうことができ，伝統的モンゴル文字文書のディジタル図書館システムの有用性を確認した．

2．伝統的モンゴル文字とモンゴル語

モンゴル国，中国及びロシア連邦の複数の地域に住んでいるモンゴル人の間ではモンゴル語が使用されている．モンゴル語はアルタイ言語の一つである．モンゴル語には独特の伝統的モンゴル文字（以下モンゴル文字と表記）とキリル文字の現代モンゴル文字（以下キリル文字と表記）を使用したモンゴル語の 2 つのスクリプトがある．

モンゴル文字の起源はセム系文字の一種であるアラム文字になる．アラム文字からソグド文字が形成され，8 世紀ごろ，ソグド文字からウィグル文字が形成された[2].

13 世紀にチンギス・ハーンがモンゴル帝国を築いたとき，ウィグル文字を借りてモンゴル語を表した．モンゴル族は Phags-pa，Todo および Soyombo のようないくつかの書記体系を作り，使用してきた．

ブリヤート共和国のモンゴル族・ブリヤート人はブリヤート方言から作ったブリヤート文字を使っていた．カリムイク共和国のモンゴル族・オイラート人は Todo 文字を使っていた．モンゴル文字のもう一つの派生文字は満州文字である．

本研究ではモンゴル文字のみを扱う．モンゴル文字は表音文字である．モンゴル文字は縦に書かれ，日本語と違い，左から右に行が進行する．モンゴル文字の字は単語中の位置によって異なる形をとる．モンゴル文字の文字集合は子音文字が 27 文字，母音文字が 8 文字で合計 35 文字からなる．

モンゴル国ではキリル文字が公式に使用される文字になって 60 年が経つ．ロシア語のキリル文字集合に，モンゴル語固有の母音をあらわす ө と ү の 2 文字が追加されている．文字集合は母音文字 13 文字，子音文字 20 文字，記号 2 文字の 35 文字からなる．語順は日本語と同じく，主語－補語－述語の SOV である．

近年，伝統的モンゴル文字と現代モンゴル語の話し言葉が変化してきている．したがって，検索においては現代モンゴル語からモンゴル文字への変換が重要である[3].

3．先行研究

我々はモンゴル文字のディジタル図書館の構築に関する研究において現代モンゴル語のキーワードでモンゴル文字のテキストを検索するシステムを提案している[3][4]．この論文では，モンゴル文字テキストの表示および Unicode における変換方法について説明している．これにより，Unicode 規格に登録されたモンゴル文字が活用できる．

また，筑波大学の満らにより，モンゴル文字情報処理に関する研究が行われている[5]．伝統的モンゴル語と現代モンゴル語の表記規則に基づいた文字単位で相互変換する翻字手法を提案している．まず，原言語テキストから単語抽出，助詞処理，長母音処理を行い，文字変換を正字法に適用し目的言語テキストを出力するような複雑な処理を行っている．この研究では，ローマ字転写による電子化方式を採用しており，現代モンゴル語による検索には対応していない．

4．モンゴル文字の情報処理に関する問題点

モンゴル文字の情報処理においては，その特徴から以下の問題が挙げられる．

現在のところ，Microsoft Office 2007 で使用できるモンゴル文字の IME が出ているが，制御記号が使用されてない．また，モンゴル文字に対応した一般向けの Web ブラウザは存在しない．また，モンゴル文字は縦書きで，行が左から右に進行する．現在の HyperText Markup Language（HTML）および Cascading Style Sheets（CSS）は右から左に行が進行する言語には対応しているが，モンゴル文字には対応していない．

モンゴル国でキリル文字が導入されて以来，モンゴル文字と現代モンゴル語の話し言葉が変化してきており，モンゴル語の情報処理に大きな問題を引き起こしている．中国内蒙古自治区の研究者らがモンゴル文字に関する多数の研究を行っている．モンゴル文字の古文書のディジタル図書館を構築するためには，モンゴル文字テキストと現代モンゴル語のテキストを結びつける方法が必要である．また，現在使われているモンゴル文字のフォントには異型字と結合字が含まれてないため，現在の Unicode フォントで（Mongol_Script.tff）はモンゴル文字の正しい表示には不十分である．

5．モンゴル文字ディジタル図書館

5．1 システムの概要

本節では，キリル文字の入力によってモンゴル文字の文書の検索を可能にするディジタル図書館の構成について述べる（図 1）．入力に関して，キリル文字だけではなくラテン文字を用いることも考えられる．しかし，現在キリル文字が公式に使われているため，利用者にとってキリル文字が最も使いやすいと考えられる．本システムではキリル文字を採用しているが，ラ

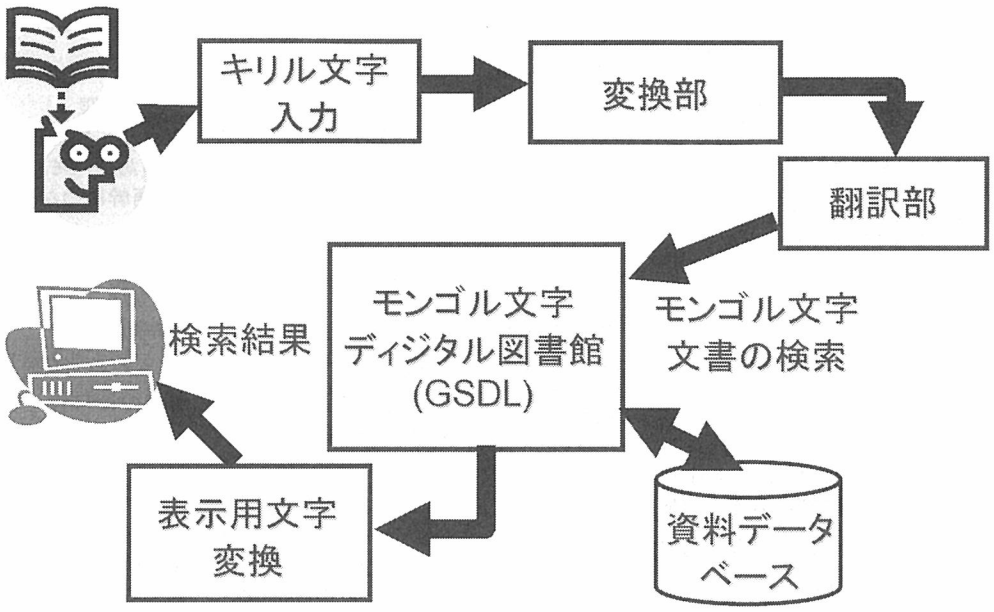

図 1: モンゴル文字ディジタル図書館の構成

テン文字に対応することも容易である．

本研究ではニュージーランドの Waikato 大学で開発された Greenstone ディジタル図書館(GSDL)システムを使用する．Greenstone システムは Unicode に対応した，ディジタル図書館コレクションの構築や配布のためのオープンソースソフトウェアである[6]．このソフトウェアはディジタル図書館システムとして現在最も普及しているものである．

本研究では，利用者が最も利用しやすい環境を考慮した検索手法を提案する．2 章で説明したようなモンゴル文字の特徴を踏まえ，3 章で述べたモンゴル文字情報処理に関する過去の研究と異なる手法によるディジタル図書館システムを提案する．提案するディジタル図書館システムでは，利用者がキリル文字の検索キーワードを入力すると，変換部により現代モンゴル語の検索質問に変換され，翻訳部に送られる．ここで，検索質問を単語に分割し，辞書により伝統的モンゴル文字の単語に変換する．これを

用いて Greenstone システムで検索が行われ，最終的に伝統的モンゴル文字文書の検索結果が表示される．これらの変換部及び翻訳部について 5.2 節と 5.3 節で説明する．

5. 2 変換部

ディジタル図書館において，検索エンジンは最も重要な構成の一つである．本システムでは，既存の IME (Input Method Editor)を利用し，現代モンゴル語のキーワードを入力する．IME とは，Windows システム上で日本語や中国語など，文字の多い言語で入力を行うために必要な変換ソフトウェアである．利用者はキリル文字もしくはラテン文字のアルファベットでキーワードを入力する．

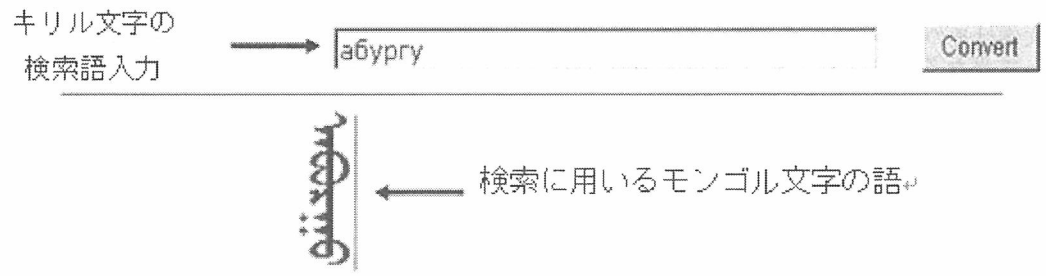

図 2: 伝統的モンゴル文字の変換インタフェース

例えば，「аврага」（「優勝者」）と発音され，モンゴル文字で「абурги」書かれる単語を入力すると，「абурги」を含んでいる文書を検索し，それが正しいかを調べ，データベースからこの単語を含んでいるすべての文書を検索結果画面に縦書きで表示する．検索語変換の例を図 2 に示す．現在は使用していないが，今後この処理において現代モンゴル語とモンゴル文字の辞書を使用する予定である．

5．3 翻訳部

ここでは，変換部で処理できない複雑な処理が行われる．2 章で述べたようにモンゴル文字の字は単語中で異なる形をとる．モンゴル文字の正字法に基づく変換条件を以下に挙げる[7].

(1) 単語中の位置
単語中の位置によって文字の形が変わる．「ᠠ」（a）は語頭に入ると「ᠠ」，語中に入ると「ᠷ」，語尾に入ると「ᡝ」の形をとる．

(2) 音節中の位置
文字は単語の同じ位置にあっても，その音節の中の位置が異なり，したがって異なる形をとる．モンゴル文字の「ᠨ」（н）子音は一つの音節の中に入ると語頭なら「ᠨ」，音節の終わりに入ると「ᡱ」の形になる．

(3) 音節中の単語中の位置
文字は語中に入るとその音節が単語のどの位置にあるかによって異なる形をとる．モンゴル文字の「ᠤᠷ」（γ）という母音は同じ語の中に入るとその音節が語頭の場合「ᠤᠷ」の形，語中の場合「ᡠᠷ」の形をとる．

(4) 前の文字の制限
文字はその文字の前に書かれた文字によって異なる形をとる．モンゴル文字の「ᡰ」（r）文字は，「ᠰ」（c）と「ᠳ」（д）の後ろに入ると「ᡳ」ではなく，「ᡳ」の形をとる．

(5) 母音の発音の特徴
文字はその文字が入った単語の母音の発音によって異なる形をとる．モンゴル文字の「ᡰ」（r）は大声の発音の母音の場合「ᡳ」，空の発音の場合「ᡵ」になる．

(6) 2 つの語根
モンゴル文字の場合，特に人名，地名は 2 つの語根から成る．そのため，2 番目の語根の語頭文字は特別な形をとる．「ᠮᠣᡰᡳᡵᡝ」のような単語の 2 番目の「ᠠ」（a）は「ᠷ」ではなく「ᡳ」である．

(7) 結合・切り分けた書き方の制限
文字は語中で結合する書き形や，独立形とその独立グリフで書かれる場合がある．モンゴル文字の「ᠺ」（и）は，単語の語幹につなぐ「ᡳ・ᠷ・ᡵ」のような形を持っている．また，切り分けて書く「ᡵ」，独立グリフ「ᠺ」がある．

(8) 前後の文字の書き方の特徴制限
「ᠪᠭ・ᠪᠭ・ᠴᠭ・ᠪ・ᠪ」のような弓形子音文字で始まる音節がある．その音節中の母音と子音は基本形と異なっている．基本形でつなぐと「ᠪᠭ・ᡵᠭ」の形になる．

(9) 単語の意味の使い分け
モンゴル語には，発音が似ている単語のある文字を異なる形で書き，意味を使い分ける場合がある．たとえば「ᠬᠦᠮᠦᠨ・ ᠬᠦᠮᠡᠨ」等である．

(10) 時代や書者の特徴
モンゴル文字の「ᠬᠠᠯᠠᡰᡠᡳᡵᡝ」（halagun）と書く単語は，昔「ᡵᠠᠯᠠᡰᡠᡳᡵᡝ」と書かれていた．

以上（1）～（10）をモンゴル文字の基本変換条件としてコード変換に用いる．

5．4 表示用文字変換

Unicode 規格には基本文字セット，句読記号，数字が登録されている[8]．Unicode 規格に含まれない文字形を含む文字コードも存在するが，索引付けや検索が複雑になってしまうため，本システムでは Unicode の基本文字で文書を保存し，表示の際には Unicode の私用領域を使用した Mongol_Script.ttf というフリーのフォントを用いて異形字や結合字を表示する[9].

Unicode では，符号化される単語の曖昧性を解決するために制御記号が使用される．この制御記号は自由選択記号“FVS”(Free Variation Selector)と“MVS”(Mongolian Vowel Separator)からなる．“FVS”の使用例を表 1 に示す．

表 1：“FVS”(Free Variation Selector)の使用例

文字の組み合わせ	表示
ᡵ FVS1 ᠷ	ᡵ
ᡵ FVS1 ᠨ	ᡵ

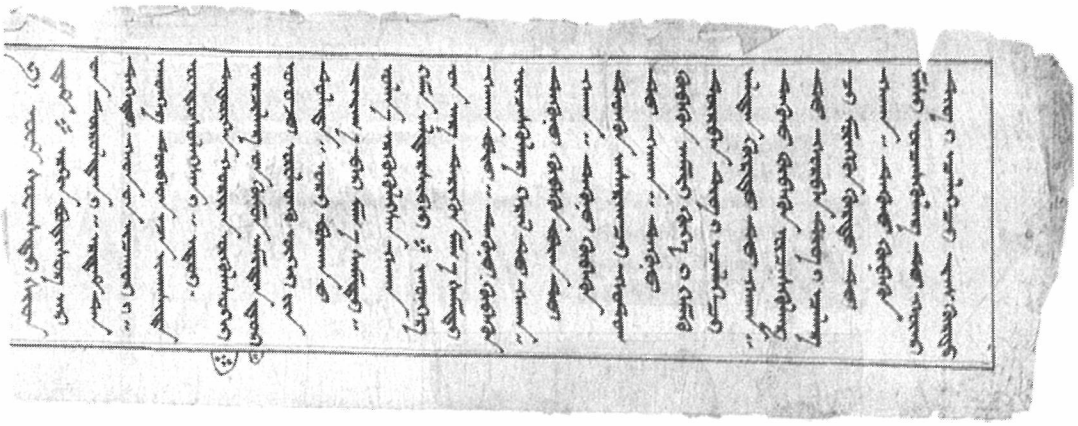

図 3: 「アルタン・トブチ」のページ

“MVS”は母音が独立形をとる場合に使用される.

制御記号と基本文字を使用することで，正しくモンゴル文字を表示することができる．これはモンゴル文字ディジタル図書館を実現する上で非常に重要である.

6. Greenstone システムへの実装

本研究では，実験対象に「アルタン・トブチ」という歴史書を用いる[10]．この書籍は遊牧民のもとで保存されてきて，1997 年に研究者が発見し，その後研究され始めた．いくつかの種類が存在する中で，モンゴルに残されている唯一の歴史書である.

この書籍は 1604～1628 年の間に書かれたと推定されている．チンギス・ハーンとモンゴル帝国を書いた「モンゴル秘史」（元朝秘史）の次のモンゴル史に関する貴重な資料である.

「アルタン・トブチ」に使用されている言葉と句をみると，筆記や口頭史，伝説であったことがわかる．時代的には数百年さかのぼるため，語彙や文法が古代と近代の形式のものが混在する．またこの原稿は他の種類より完全であると認められている.

「アルタン・トブチ」に書かれたモンゴル史は時代的にモンゴル帝国時代（XIII～XIV）と政界分裂時代（XV～XVII）に大きく分けられる．特に，政界分裂時代である XV～XVII 世紀ごろのモンゴル史が適切かつ明確に書かれていることが歴史研究に重要である.

この書籍の体裁は，37×7.5 センチサイズの少々厚めの紙に，黒と赤の墨を使い，30×5.5 センチのフレームの中に 29 行で書かれた，164 枚の経典式の本である.

この書籍は，モンゴルの他にロシアや中国で歴史，言語学的に研究されている．字の形がほとんど変わらず，竹のペンで丁寧に書かれている（図 3）.

Unicode 規格に対応した伝統的モンゴル文字の IME が未だに存在しないため，モンゴル文字ディジタル図書館に用いる「アルタン・トブチ」のテキスト作成は人手で行い，多くの時間と労力を費やした.

モンゴル文字コレクションを作成する際に，Greenstone システムの基本ソースコードやマクロファイルは一切変更せず，自らが作成した機能やマクロファイルをそれぞれ実装した．マクロファイルは主に，キリル文字からモンゴル文字に変換する部分および 5.3 節で説明した基本変換条件を調べる 2 つの部分からなる.

キリル文字からモンゴル文字への変換は文字単位で行う．モンゴル語は分かち書きされるため，空白によって単語が抽出される．コレクションの一部を図 4 に示す．現在のところキリル文字入力からモンゴル文字検索語に変換し，検索を行っている．図 5 に変換されたモンゴル文字の検索語及び検索結果を示す．現在，検索に用いる単語の選択が一つになっているが，モンゴル文字の辞書が使用できるようになった際には複数の単語の中から選ぶことができるようにする予定である．図 6 に検索された文書の例を示す.

7. 実験

本節では，モンゴル文字文書をキリル文字で検索する手法を用いる実験と「アルタン・トブチ」のディジタル図書館の利便性について検証するために行った利用者評価実験の結果について述べる.

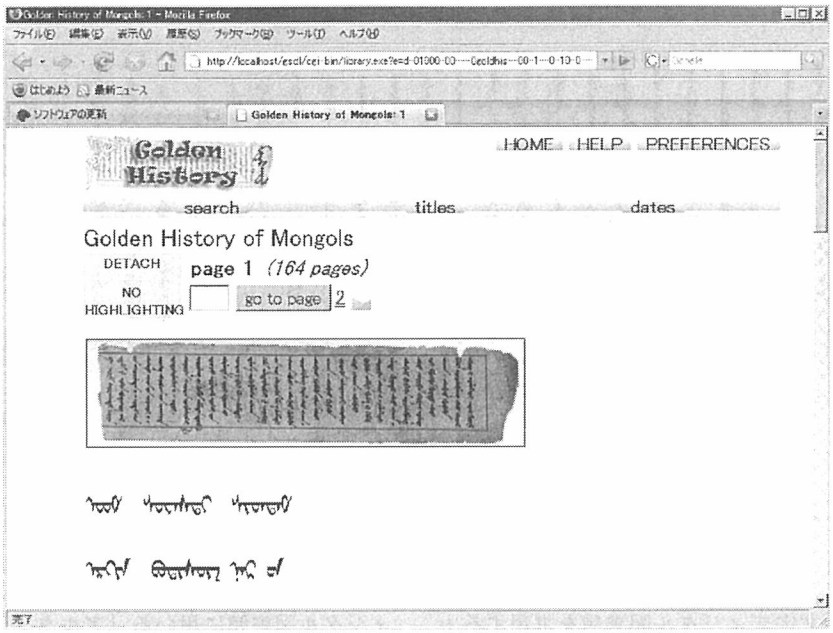

図 4: モンゴル文字文書の Greenstone コレクション

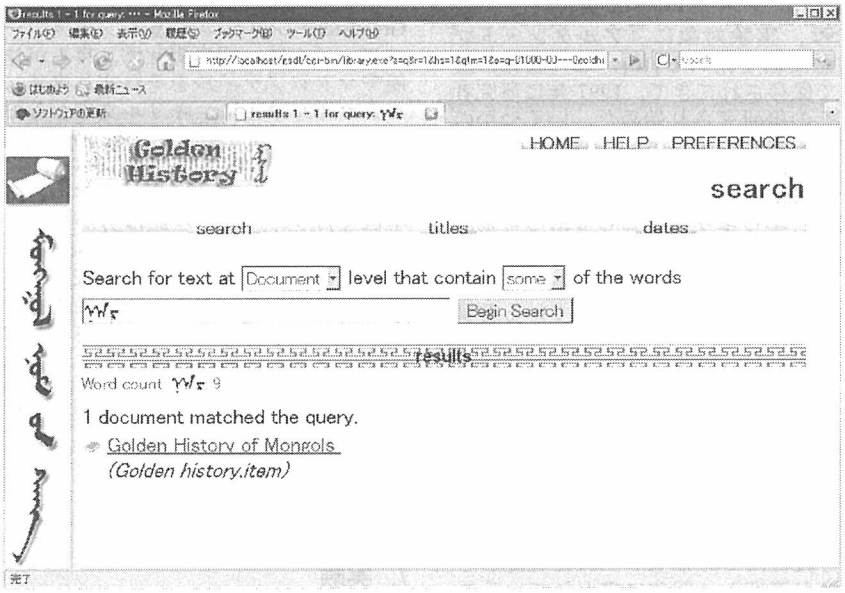

図 5: 検索結果画面

7. 1 検索実験

「アルタン・トブチ」のコレクションに対してキリル文字の検索語を用いた文書検索の実験について述べる.

5.3 節で説明した基本変換条件にしたがい, 主に名詞を用いて検索を行う. 検索を行う単語抽出などの処理は Greenstone システムのデフォルト設定機能で対応している. モンゴル語の単語は分かち書きされるため, 空白によって単語が抽出される.

最も基本規則になる変換条件を調べ, キリル文字による検索語入力を用い, 検索を行った. そして, 「アルタン・トブチ」の文学研究書に掲

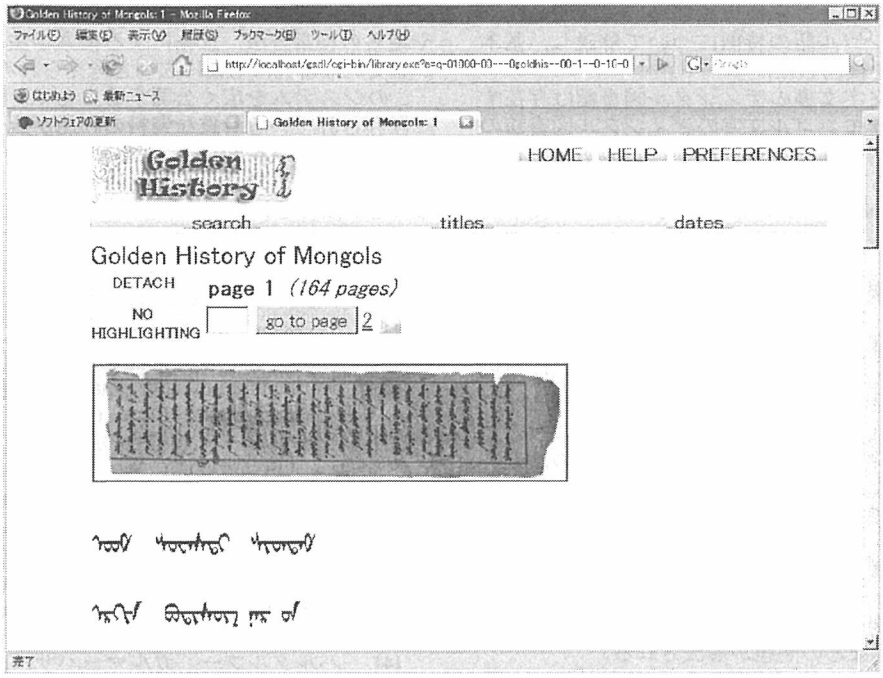

図 6: 検索された文書

載されている単語リストと検索された単語回数を比較した．全文書中に 20 回以上現れる単語が 127 個ある．その中から名詞と数字名を用いて行った検索語の一部を表 2 に示す．

実験の結果，検索出現回数は，「アルタン・トブチ」の文学研究書に掲載された単語のリストと同じ回数となり，本研究で取り上げたキリル文字からモンゴル文字に変換する基本変換条件をほぼ正しく行えることを示すことができた．

今後，モンゴル文字の辞書が入手できれば，曖昧性が解消され，他の正字法や不規則な単語を用いた検索が可能になる．

7. 2 利用者評価実験

本節では，モンゴル文字ディジタル図書館における検索手法の有用性を検証するために行った利用者評価実験について述べる．対象として本研究で提案したモンゴル文字ディジタル図書館である「アルタン・トブチ」を用いる．

表 2：実験に用いた検索語例

キリル文字	モンゴル文字（日本語意味）	回数
Эзэн	 （主人）	146
Жил	 （年）	86
Энэ	 （この）	86
Зарлиг	 （命令）	65

モンゴル語を母国語とするモンゴル人の被験者 5 名において利用者評価実験を行った．この実験はモンゴル文字の他の変換手法に比べ，どのような利点があるかを利用者に評価してもらい，システムの有用性を調べるための実験である．

この実験は，現在の公用文字であるキリル文字の検索語を用いてモンゴル文字で書かれた文書を検索する手法にどのような利点があるかについて，客観的な立場から利用者に評価してもらう．利用者自身が求めている結果であったかを満足度で表し，5 段階で評価してもらった．

表 3 の結果より，どの項目においても検索手法およびシステムが高く評価されていることがわかる．また，モンゴル文字文書のディジタル図書館システムを利用した感想を聞くと，「現在アクセスできなくなっている古文書を利用することができた」「画像の表示が大変よくできているので現物を見たようでよかった」などの良い意見が出された．しかし，文書が横に表示されている点については評価が比較的低かった．

8. まとめ

本研究では，モンゴル文字文書のディジタル図書館の構築について述べ，利用者実験を行った．現代モンゴル語のキーワードでモンゴル文字のテキストを検索する時代横断型の検索システムを提案し，それに必要なモンゴル文字テキ

ストの表示および変換方法について説明した．また，モンゴル語の特徴について解説し，基本変換条件を検討した．

キリル文字文書のディジタル図書館は存在するが，近年モンゴル文字のディジタル図書館の利用ニーズが高くなっており，実装が強く要求されている．

Greenstone システムにおけるモンゴル文字文書のディジタル図書館を実装し，モンゴル文字の表示や，キリル文字からモンゴル文字に変換する基本変換条件に対応した機能が正しく動作している．しかし，現状では検索語として名詞のみしか扱えない．

提案したシステムの利用により，Unicode 規格に登録されたモンゴル文字の利用が活発になると考えられる． ただし，Unicode 規格に対応したモンゴル文字の IME が存在しないため，モンゴル文字文書のディジタル図書館のテキスト作成を人手で行う必要があり，相当の時間と労力がかかる．今後の課題の一つとして，モンゴル文字テキストの入力方法の開発が挙げられる．また，モンゴル文字の辞書が実現できれば，モンゴル文字の他の正字法や不規則な単語に適用し，曖昧性などの問題が解決される．

実験では基本変換条件にしたがうキリル文字検索語による検索ができることを確認した．利用者実験では，キリル文字入力を使用し検索を行う手法が大変良かったという点からシステムとしては高い評価が得られた．

Greenstone システムの最新の技術であるページめくり(page turning)機能を適用することも考えられる．モンゴル文字の経典は，ページめくりが普通の本と異なり，下から上へとページめくりがされる．また，この研究ではモンゴル文字のみを扱っているが，モンゴル文字から形成された他のスクリプトにも適用可能かどうか検討が必要である．

表 3：利用者評価の結果

評価項目	評価者					平均
	A	B	C	D	E	
キリル文字のキーワードによる検索	5	4	5	4	4	4.4
モンゴル文字文書を検索数する際にこの手法の有用性	4	4	5	4	4	4.2
全体的に表示された画像の表示	5	5	5	5	5	5
全体的に表示された文書の表示	4	4	4	4	3	3.8
他の手法と比較した有用性	5	4	5	4	4	4.4
システムの全体の評価	5	5	4	5	4	4.6

さらにモンゴル文字の歴史書に登場する人名や地名の検索方法，意味による検索手法の実現が歴史的な研究にも役に立つと考えられる．

このシステムを広く公開することで，図書館に保存されている貴重な資料の利用が活性化でき，またモンゴル国だけではなく世界のさまざまの国にいるモンゴル人が利用できるようになることが期待される．

参考文献

[1] Тунгалаг, Д.: Монгол улсын үндэсний номын сан дахь монголын түүхийн гар бичмэлийн номзүйн судалгаа, 1-р боть, Тайм принтинг, 2005. (モンゴル語)

[2] 三上善貴: 文字符号の歴史—アジア編—, 共立, 2002.

[3] Garmaabazar, Kh. , Maeda, A.: Retrieval Technique with the Modern Mongolian Query on Traditional Mongolian Text, In Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL2006) , pp.478-481, 2006.

[4] ハルタルフー ガルマーバザル，前田亮：伝統的モンゴル文字文書のディジタル図書館の構築，人文科学とコンピュータシンポジウム（じんもんこん2006）論文集，pp.319-326，2006.

[5] 満都拉，藤井敦，石川徹也: 伝統的モンゴル語の電子化方式とテキスト検索への応用, The IEICE Transactions, D-II, Vol. J88-D-II, No.10, pp.2102-2111, 2005.

[6] http://www.greenstone.org/

[7] Насан-Урт, С.: Монгол хэл бичгийн сураг занги боловсруулах онол практикийн зарим асуудал, 2004. (モンゴル語)

[8] The Unicode Consortium: The Unicode Standard 4.0. Addison-Wesley, Boston San Francisco New York, 2003.

[9] Erdenechimeg, M., Moore, R., M., Namsrai, Yu.: UNU/IIST Technical Report No.170 – Traditional Mongolian Script in the ISO/Unicode Standards, 1999.

[10] Чоймаа, С.: Хаадын хураангуй алтан товч, 2002. (モンゴル語)