

(イ) 固有名表示

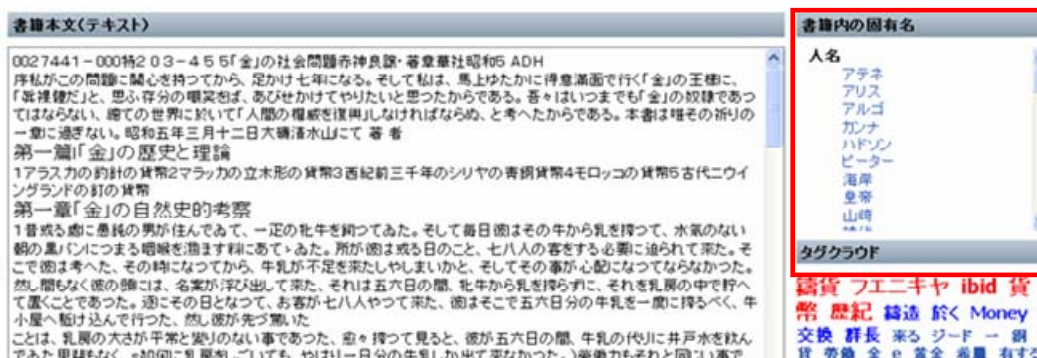


図 2-19 検索表示システムの固有名表示

固有名表示とは、書籍の本文中に出現する固有名表現を抽出し、人物名・地名等に分類して表示するものである。典拠データ¹⁰および Wikipedia の見出し語を元に固有名を抽出した。本文の中に含まれる固有名を表示することで、内容を類推するための手がかりとすることを企図した機能である。

(ウ) タグクラウド表示

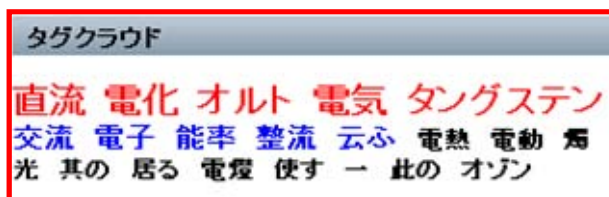


図 2-20 検索表示システムのタグクラウド表示

タグクラウド表示とは、書籍の本文から特徴語とスコアを抽出し、特徴語をスコアが高い順にフォントのサイズを大きく、色付けして表示する機能である。利用者が書籍の内容を類推するための手がかりをわかりやすく提示するための機能である。連想検索エンジン GETAssoc を利用して実現した。

(d) 本文表示画面

本文表示画面の概要を以下に示す。

¹⁰ 書誌的記録（書誌レコード）の標目となる個人名、団体名、統一タイトル、シリーズ名、件名などの典拠形を定めたデータ。



図 2-21 本文表示画面

(ア) 目次・本文リンク



図 2-22 検索表示システムの目次・本文リンク

目次・本文リンクとは、目次に設定されたリンクをクリックすることにより、本文の該当箇所へ遷移できる機能である。

(イ) 検索語出現数表示

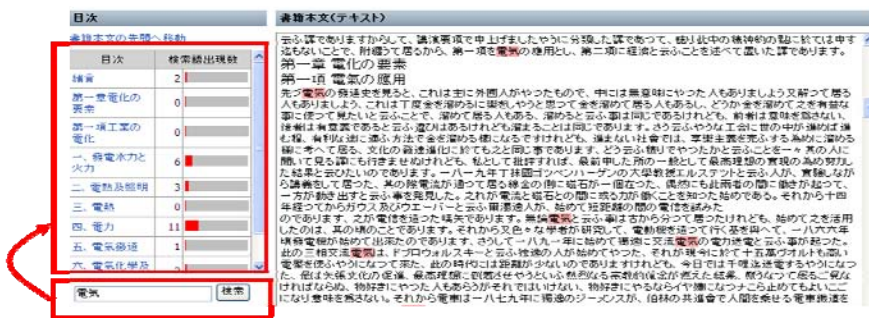


図 2-23 検索表示システムの検索語出現数表示

検索語出現数表示とは、目次の各章ごとに検索語の出現数を表示するものである。書籍の中で検索語と関連の強い記述部分を把握するための手がかりを提供するための機能である。

(ウ) 検索語ハイライト表示

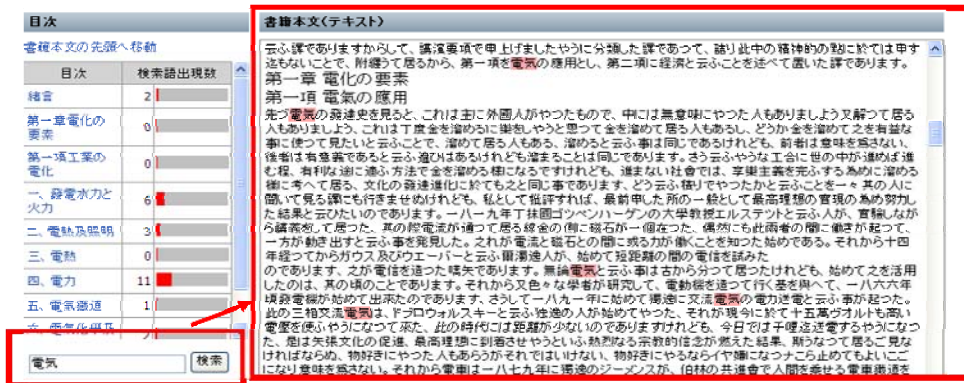


図 2-2 4 検索表示システムの検索語ハイライト表示 (テキスト表示)

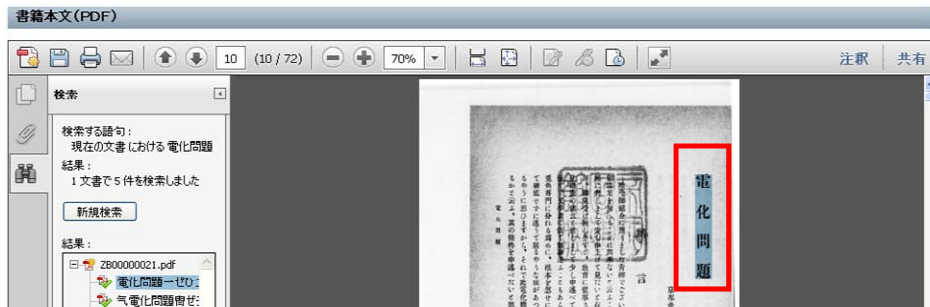


図 2-2 5 検索表示システムの検索語ハイライト表示 (PDF 表示)

検索語ハイライト表示とは、本文中の検索語に対して、背景の色を変えて表示するものである。本文の中から検索語の位置を容易に把握できるようにするための機能である。

2. 3 実証実験の対象書籍

本実証実験で対象とした書籍を以下に示す。

(1) テキストデータ作成に関する実証実験の対象書籍

本実証実験では、国立国会図書館が所蔵する、以下の書籍の全文テキストデータを作成した。

表 2-1 全文テキストデータ作成対象書籍（国立国会図書館所蔵資料）

年代 ID	全文 ID ¹¹	書籍名	処理対象	
			書籍の一部 / 全頁	処理 ページ数
明治 01	07	一元哲学：最新	一部	60
明治 02	06	哲学概論	一部	58
明治 03	81	オセロー	一部	60
明治 04	82	至極重宝 一名・東京案内	一部	60
明治 05	70	噫無情	一部	60
明治 06	71	和仏法律学校講義録. 雑報 36 年度 第 3 部	全頁	112
明治 07	72	初等代数学	一部	60
明治 08	80	四書白文	一部	60
明治 09	73	東京印刷同業組合名鑑並ニ材料商名鑑	一部	60
明治 10	74	東京開成学校自火及近火消防	全頁	16
明治 11	75	育嬰草	一部	60
明治 12	76	飲水要論	一部	60
大正 01	18	作家の日記. 上巻	一部	60
大正 02	02	電化問題	全頁	72
大正 03	01	能率的販賣經營法：人間學應用	一部	102
大正 04	21	内国保険会社信用録. 第 1 回	一部	60
大正 05	20	エスペラント語速成教科書：短期講習用	全頁	62
大正 06	77	在米日本人人名辞典	一部	76
大正 07	78	政治学研究	一部	60
大正 08	79	変態心理学講義録. 第 1 篇	一部	60
大正 09	22	岩崎弥太郎	一部	60
大正 10	23	牛乳の飲み方	一部	60
昭和戦前 01	05	企業整備令と小売業整備要綱	全頁	350
昭和戦前 02	24	受験全書行刑法編：附・会計法編	一部	60
昭和戦前 03	25	無産者法律必携	一部	60
昭和戦前 04	26	第五十六回帝国議会大演説集	一部	60
昭和戦前 05	27	営団経済の倫理	一部	60
昭和戦前 06	28	栄えゆく道	一部	74
昭和戦前 07	29	強制執行並保全処分概論	一部	60
昭和戦前 08	30	無尽・保険・月賦・其他の掛金にたいする法律戦術	一部	60
昭和戦前 09	31	「金」の社会問題	一部	51
昭和戦前 10	32	憲法論	一部	60
昭和戦後 01	17	初等統計解析	一部	74
昭和戦後 02	08	梶井基次郎全集. 第 1 巻	一部	104
昭和戦後 03	12	戦国人名事典	一部	60
昭和戦後 04	11	鎌倉・室町人名事典：コンパクト版	一部	58
昭和戦後 05	14	マルクス=エンゲルス全集. 第 18 巻	一部	60
昭和戦後 06	16	資料日本現代史. 2	一部	52
昭和戦後 07	09	マルクス資本論草稿集. 7	一部	120
昭和戦後 08	10	日本の名著：近代の思想	一部	60
昭和戦後 09	13	世界の名著：マギアヴェリからサルトルまで	一部	60
昭和戦後 10	15	宦官：側近政治の構造	一部	60
雑誌 01	19	文芸春秋	一部	68
合計				3,069

これらの書籍のうち、明治～昭和戦前の書籍は、2 値モノクロ¹²またはグレイの画

¹¹ テキスト化システムで処理した順に付番した番号。後述する評価に用いる。

¹² 完全な白色と完全な黒色の 2 色のみで表現された画像のこと。

像データであり、以下のような特徴を持つものを含む。

- ・文章中に縦中横¹³、数式、漢文、割注¹⁴がある
- ・デザインの異なる活字が混在している
- ・漢字と仮名サイズが混在している
- ・左横書きおよび日英字が混在している
- ・傍点やルビがある

また、昭和戦後の書籍は、以下のような特徴を持つものを含む。

- ・多段組みである
- ・索引がある

上記の書籍に加えて、出版社 25 社から提供された書籍の電子ファイルの一部に対しても、構造化作業を実施し、全文テキストデータを作成した。

表 2-2 全文テキストデータ作成対象書籍（出版社提供データ）の概要

区分	タイトル	冊数	ページ数	フォーマット (冊数)			
				PDF	TXT	.book	XMDF
図書	326	326	62,042	165	88	15	58
雑誌	1	20	40	20	0	0	0
総計	327	346	62,082	185	88	15	58

(2) 全文テキストデータの検索・表示に関する実証実験の対象書籍

本実証実験では、「テキストデータ作成に関する実証実験」で全文テキストデータを作成した書籍に加えて、OCR で読み取り、校正・構造化を行っていないテキストデータ 19,621 件と、そのほかに書誌データ 382,000 件を蓄積して評価を行った。

¹³ 文字のレイアウト設定に用いられる機能のひとつで、縦書きの文書に半角文字（数字やアルファベット）を横並びの状態を組み込む機能のこと。

¹⁴ 任意の文字列を小さくし、一行の中に二段構えで表示させる機能のこと。