インターネットを用いた研究支援環境
～情報検索システム～

A Research Support System Using the Internet
— An Information Retrieval System —

石田 崇 †*　足立 鉱史 †**　後藤 正幸 ‡　酒井 哲也 §　平澤 茂一 §§
Takashi Ishida†*　Hiroshi Adachi†**　Masayuki Gotoh‡　Tetsuya Sakai§　Shigeichi Hirasawa§§

†* 早稲田大学 大学院理工学研究科 (現 早稲田大学 理工学部経営システム工学科)
† * * 早稲田大学 大学院理工学研究科 (現 新日鉄ソリューションズ (株))
‡ 武蔵工業大学 環境情報学部,　§(株) 東芝 研究開発センター
§§ 早稲田大学 理工学部経営システム工学科

† Graduate School of Science and Engineering, Waseda University
‡ Faculty of Environment and Information Studies, Musashi Institute of Technology
§Knowledge Media Labolatory, Research and Development Center, Toshiba Co., Ltd.
§§School of Science and Engineering, Waseda University

要旨： インターネットを用いてセミナー・研究発表・討論などを行うための研究支援環境を構築する．このシステムは (1) 電子会議システム，(2) データベースシステムからなる．前者はいわゆるエレクトロニックカンファレンスシステムであるが，ランニングコストを抑えるために専用線を避けインターネットを用いている．後者はセミナーなどで用いる資料・データ・論文などを検索し，効率よく議論するするためのものである．いずれも低価格な汎用機器を利用して実現することを考えている．本稿では後者の文書検索システムに焦点を当てる．情報検索システムの導入・運用の準備段階として，実際に研究資料データベースを構築し，検索精度の検証を行ったのでその結果を報告する．

Abstract: A research activities support system for researchers over worldwide universities and industries is developed. Because of reducing the running cost, we construct (1) network conference system by using low cost devices and the Internet. Besides (1), we introduce (2) a private database system and its information retrieval system which are installed on usual low cost personal computers (PC's). Based on these subsystems, research activities for researchers such as discussions with each other over the countries can be supported. In this paper, we introduce our private information retrieval system. And we show the results of the experiment of precision of the system.

1　はじめに

最近の情報通信システムの発展にともない，教育支援システムが広く普及しつつある．その一方，古くからネットワークを用いた研究者同士の情報交換が行われていたにもかかわらず，研究支援システムはさほど身近に実用化されているとは言い難い．本報告では空間的に離れた大学・企業間の研究者達による共同研究の場を提供する研究活動支援システム（「Net-semi」と略称する）[1][2][3][4][5] について述べる[1].

「Net-semi」は (1) ネットワーク型カンファレンスシステム「Net-con」と (2) 研究支援用プライベートデータベース「PDB」から成っている．いずれも低価格な汎用機器を用いて研究者グループ間をインターネットで結び，学術論文や研究成果などの各種ドキュメントを共有して，情報を検索・参照しながら少人数のセミナー・研究発表・フォーラム・技術会議・講演会・技術打ち合わせ・研究指導など（以下，これらを総称して「ゼミ」という）を効率よく実施できる共同研究活動支援環境を提供しようとするものである．

本稿では「Net-semi」におけるプライベートデータベース「PDB」に焦点を当てる．「PDB」を本格的に構築・運用するにあたり，準備段階として留意点や問題点を明らかにするために，実際に研究資料データベースを構築して検索精度の検証を行ったのでその結果について報告する．まず始めに PDB の構成とデータベース構築方法について述べ，次にジャストシステム社製の検索エンジンである Concept Base Search(CBS)[6][7] による研究資料の検索実験の結果についてまとめる．また，検索エンジンを独自に設計することを目指し，ベクトル空間モデルに基づく擬似適合性フィードバック手法についても検証を行う．

[1]本システムは当初，著者らの一部が在外研究で海外に長期滞在した際の遠隔地からの研究指導用として検討を開始した．

2　研究活動支援システム「Net-semi」の構成の概要

研究活動支援システム「Net-semi」は図 1 に示すように (1) ネットワーク型カンファレンスシステム「Net-con」端末と (2) プライベートデータベース「PDB」検索システムで構成される．これにより共通の興味を持つ大学や企業の研究者は物理的に離れていても，インターネットを通して共同研究を実施することができる．

2.1　ネットワーク型カンファレンスシステム「Net-con」

「Net-con」の目的は従来の電子会議システムと基本的に同じであるが，コストや運用上の観点から専用線などの高速回線ではなく，インターネット回線を用いることを前提とし，また，全体的に比較的低価格の機器によって構成されている．図 2 に「Net-con」の構成を示す．

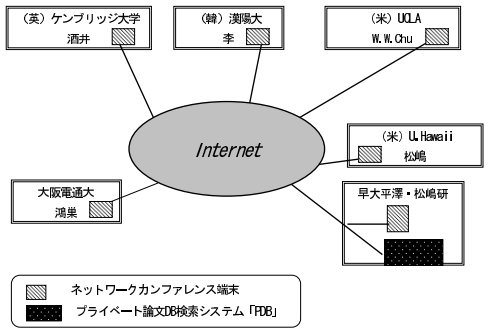

図 1: 研究支援システム「Net-semi」の構成

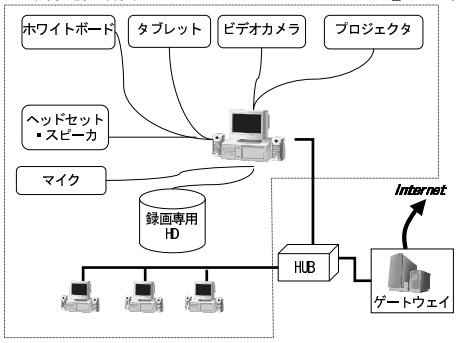

図 2: 「Net-con」の構成

発表者はスクリーンやディスプレイ上の図表や数式を指しながら声・挿絵・動画などにより研究成果を説明する．その結果，遠隔の研究者同士が容易かつ迅速に討論することができる．

通常の電子会議システムは専用線や高速回線を用いるため，常に通信サービス品質が保証されているが,「NetCon」はインターネットを用いるため，とりわけそのQoSが重要な問題である [1][2][3][4][5].

2.2 プライベートデータベース検索システム「PDB」

プライベートデータベース「PDB」は各種データを蓄積するデータベースサーバと文書検索エンジンから構成されている.「PDB」で蓄積・共有されるドキュメントは論文やレジュメ，発表資料などであり，これらはPDFファイルや，MS Word, MS Power Pointファイルなどの電子データである．さらに，用紙に出力されているものや手書きのドキュメントも含めるためデータをデジタル化するスキャナとOCRが必要となる．OCRにより資料はテキストデータを保持したPDF形式に変換してデータベースに格納される．

なお，データベースの登録は特定のクライアントからに限定されるが，検索・出力結果表示はIDとパスワードの入力により，ブラウザを通じてインターネット上の任意のクライアントから接続が可能である．図3に「PDB」の構成を示す.

3 PDBにおける情報検索システム

3.1 Concept Base Search: CBS

本システムでは検索エンジンとしてジャストシステム社のConcept Base Search (CBS) [6][7] を用いている．CBSは「概念検索方式」を採用しており，ベクトル空間モデルに基づいてドキュメント間の類似度（関連度）を目安に，ユーザが提示した検索質問に近いドキュメントをランク付けして回答する．そのため，

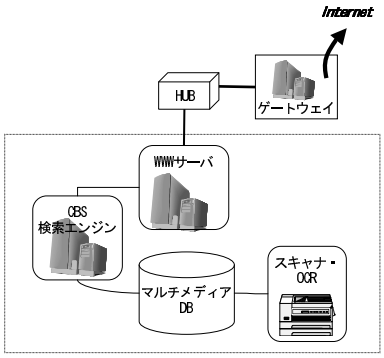

図 3: 「PDB」の構成

1. キーワード検索や全文検索とは異なり，自然文や既存の文書ファイルによる検索

2. 検索結果を反映させてさらに関連したドキュメントの抽出

3. より適切な検索条件の生成

が可能である．これにより，多数の人間によって書かれた文書の集合のように，用いられる言葉の統制が取られていないデータセットに対しても検索が有効に機能する．

また，CBSはマルチドキュメントフォーマット対応で多くの文書形式のファイルをそのままデータベースに格納できるため，登録作業が容易である．

3.2 ベクトル空間モデル (Vector Space Model:VSM)

情報検索システムは，検索対象の文書集合に検索質問を与え，各文書と検索質問との類似度を計算することにより検索結果を得る．このプロセスには幾通りかの方法が提案されており，これらを数理的に記述する情報検索モデルがある [8].

ベクトル空間モデル (VSM)[8] はもっとも一般的な情報検索モデルである．VSMによる検索システムでは，形態素解析処理により全文書から索引語 w_j を抽出し，この索引語を次元として，文書をベクトルで表現する（文書ベクトル d_i $(i = 1, 2, \cdots, I)$）．検索モデルは文書集合を表す索引語-文書行列 $A = [a_{ij}]$ と検索質問ベクトル q によって特徴付けられる．一般に $a_{ij} \geq 0$，$q_j \geq 0$ は非2値で，それぞれ文書や，検索質問に出現した索引語 w_j の重み（重要度）を用いる．多くの場合，a_{ij} は文書 d_j における索引語 w_j の出現頻度 (trem frequency:tf) と全文書中で索引語 w_j の出現した文書数の逆数 (inverse document frequency:idf) の積 (tf-idf 値) が用いられる．前者は局所的重み係数，後者は大域的重み係数である．

ベクトル空間モデルでは文書ベクトル d_j と検索質問ベクトル q 間の類似度を余弦（次式）や内積で与えることにより，検索結果を文書を類似度の降順に並べたランキングで提示することを可能に

している．

$$sim(d_j, q) = \frac{\sum_{j=1}^{J} d_{ij} \cdot q_j}{\sqrt{\sum_{j=1}^{J} {d_{ij}}^2} \sqrt{\sum_{j=1}^{J} {q_j}^2}} \quad (1)$$

3.3 適合性フィードバック

情報検索において，一度の検索でユーザが必要とする情報をすべて得られることはまれである．そこで，ユーザに初期検索結果を提示し，ユーザがその結果を見てシステムの挙動を変化させるように検索質問を調整することが考えられる．このような技術を一般に適合性フィードバックと呼ぶ．

ベクトル空間モデルに基づく適合性フィードバックにおいて，検索質問ベクトルの索引語の重みを修正する手法はいくつか提案されているが，Rocchioの手法 [9] が一般的である．これはユーザが適合と判断した文書集合の重心ベクトルと不適合と判断した文書集合のベクトルの差分を新しい検索質問ベクトルとする手法である．また，大規模な検索結果に対しては人手による適合・不適合の判断を行わず，最上位の文書を擬似的に適合文書とみなしてその文書だけフィードバックする擬似適合フィードバックと呼ばれる手法 [10] も存在する．

4 評価実験

4.1 データベース構築

「PDB」ではCBSを検索エンジンとして用いるため，すでにPDFやMS Word，MS Power Pointの形式で所有しているドキュメントについてはそのままデータベースに登録する事が可能である．一方，紙に出力された状態の資料や手書きのドキュメントは，スキャナで電子データとして取り込んだ後にOCR処理を行い，テキストデータの添付されたPDFファイルに変換してデータベースに登録する．なお，大規模な資料を効率的にデータベースに登録することを考慮して，スキャニングとOCRは自動処理を行う．したがって，OCRの際の誤認識などはあえて修正を行わず，そのままの状態でPDFファイルに変換することとした．

今回の実験における「PDB」の機器構成（表1）と「PDB」に登録されたドキュメントのデータ形式を表2に示す．

表 1: 「PDB」システム構成機器

機器	型名（メーカー）
PDB サーバ	Dell Dimension 8100（DELL）
検索エンジン	Concept Base Search (ジャストシステム)
OCR ソフト	読ん de!!ココ（A.I. soft）
スキャナ	Scan Snap (Fujitsu)

4.2 実験概要

PDB検索システムの性能評価を行う．本システムにおける検索性能は，紙データのスキャニングやOCR処理を含んでいるため，検索エンジンの性能に加えて，OCRの性能も影響してくる．また，CBSの内部に組み込まれている処理であるため検証することは不可能であるが，実際には日本語の処理である形態素解析の精度も検索性能に影響してくる．

表 2: PDB 登録データファイル形式

形　式		文書数
電子データ	PDF	154
	MS Word	168
	MS Power Point	206
OCR データ	PDF	779
	合　計	1,307

以上の問題点を踏まえて，以下の3つの評価実験を行う．

1. Adobe Acrobat の語句検索による OCR 性能評価
2. CBS による概念検索の性能評価
3. 擬似適合性フィードバック検索の性能評価

それぞれの実験における評価指標として，適合率，再現率を用いる．適合率と再現率は以下の式で与えられる．

$$\text{適合率} = \frac{\text{検索正解文書数}}{\text{検索文書数}}, \quad (2)$$

$$\text{再現率} = \frac{\text{検索正解文書数}}{\text{全正解文書数}} \quad (3)$$

情報検索問題においては，検索質問に対する絶対的な正解文書というものの判断は難しいが，この実験では，各個人の研究キーワードを検索質問としたときに，その当人の作成した研究資料を正解文書であると設定した．

4.3 Adobe Acrobat の語句検索による OCR 性能評価

ここでは，もともと電子データとして作成されたPDFデータと，OCR処理で作成されたPDFデータ（OCR文書）に対して，Adobe Acrobatの語句検索を行ってOCRによる認識性能について検証を行う．検索対象となるドキュメントは学生の氏名が記載された研究資料であり，氏名で語句の一致検索を行ったときの検索文書数を評価する．検索結果を図3に示す．

表 3: 各データセットに対する適合率

学生	全文書 (OCR+電子)	OCR データ			電子データ
		全 OCR	手書き	手書き以外	
A	0.73	0.60	0.00	0.88	1.00
B	0.95	0.79	-	0.79	1.00
C	0.87	0.70	-	0.71	1.00
D	0.76	0.67	0.14	0.79	1.00
平均	0.83	0.69	0.07	0.79	1.00

4.4 CBS による概念検索の性能評価

次に，表2で示した全文書を対象としてCBSによる概念検索の性能評価を行う．ここでは研究室の学生に自分の研究分野が特定できるようにキーワードを3～5語選んでもらい，これを検索質問としてCBSによる文書検索を行った．このとき，それぞれの学生が作成した研究資料を正解文書とみなす．代表的な学生5人（A～E）の検索質問例（表4）と検索結果（図4）を示す．

表 4: 検索質問例

学生	検索質問（研究キーワード）
A	セキュリティ，公開鍵暗号，鍵交換，コスト低減
B	サポートベクターマシン，優遇文書，適合性フィードバック
C	潜在的意味クラス，グラフィカルモデル，SAM，Naive Bayes
D	LDPC 符号，消失通信路，ガウス消去法，sum-product 復号法
E	短縮巡回符号，列削除，ループ 6

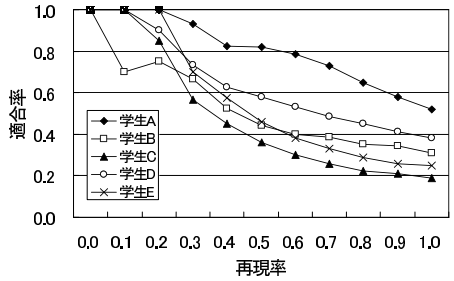

図 4: 概念検索結果（適合率・再現率曲線）

4.5 擬似適合性フィードバック検索の性能評価

「PDB」では専門的な狭い領域のドキュメントを扱っている事から，専門辞書などの利用によってより有効な検索システムを構築できる可能性がある．そこで，著者らは現在独自の擬似適合フィードバック検索エンジンの作成を試みている．CBS にも擬似適合性フィードバックと同様の機能がある事から，この性能を検証して今後の参考とする．擬似適合性フィードバックは，前述したように初期検索結果に対して最上位の文書を適合文書とみなして検索システムにフィードバックし，検索質問を拡張した上で改めて検索結果を出力する．

第 4.4 項の実験における各学生の検索質問に対して擬似適合性フィードバックを行った結果を図 5 に示す．

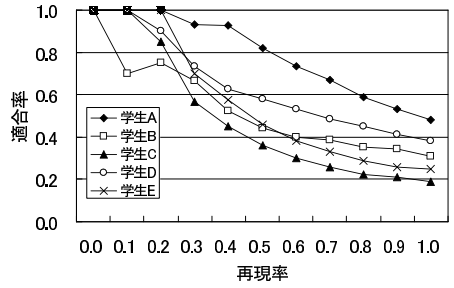

図 5: フィードバック後の検索結果（適合率・再現率曲線）

5 結果のまとめと考察

1. 表 3 より，指定の語句を含んでいる文書に対し，OCR データは概ね 70%，電子データについては 100%の割合で語句検索に引っかかることが分かった．この値が OCR の認識率であると捉えることができる．OCR データの内訳を見ると，Word や Power Point などにより電子的に作成された資料であれば紙から取り込まれた文書であっても約 80%の認識率を達成することが分かるが，一方，手書きの資料についてはほとんど認識されず役に立たないことが分かる．
2. 図 4 のグラフより，学生ごと（検索質問ごと）に適合率にばらつきがあることが分かる．これは，研究テーマそれぞれの特性と，学生が指定した検索質問の適切さにも依存する．しかし，ほとんどの学生が検索結果の上位の付近（再現率が 0 に近いあたり）で適合率が大きな値をとっており，正解文書が適切に検索されていることが分かる．
3. 図 4，図 5 より，学生 A についてはフィードバックの後に適合率の曲線がやや上昇している．これは，正解文書がフィードバックによってより検索結果の中で上位に改善された事を意味している．また，フィードバックの前後で正解文書内でも順位が大きく変化している．
4. 学生 A 以外の学生にはフィードバックによる適合率の改善が見られなかった．また，正解文書内での順位の変動も特に見られなかった．独自の検索エンジンを作成する事によって，より有効な検索を実現できる可能性がある．

6 むすび

本稿ではプライベートデータベース検索システムに焦点を当て，具体的な構築方法を示してその検索性能について検証を行った．その結果，汎用的な機器を用いてプライベートなデータベースを構築し，実用的に機能する事を示すことができた．

謝辞：本研究を行うにあたり，実験にご協力いただいた早稲田大学理工学部 平澤研究室の学生の皆さんに感謝いたします．本研究の一部は早稲田大学特定研究課題 2004A-174 の助成による．

参考文献

[1] 平澤茂一，松嶋敏泰，鴻巣敏之，酒井哲也，中澤真，李相協，野村亮，“「インターネットを用いた研究活動支援システム」システム構成”，2001 年 PC カンファレンス予稿集，pp.60-61，金沢，2001 年 8 月．

[2] 野村亮，中澤真，鴻巣敏之，松嶋敏泰，平澤茂一，“「インターネットを用いた研究活動支援システム」システム構成と評価”，2001 年日本経営工学会秋季発表大会予稿集，pp.252-253, 福岡，2001 年 11 月．

[3] 野村亮，中澤真，松嶋敏泰，平澤茂一，“「インターネットを用いたゼミと研究指導」実用化報告”，2002 年 PC カンファレンス予稿集，pp.60-61, 東京，2002 年 8 月．

[4] 中澤真，野村亮，鴻巣敏之，松嶋敏泰，平澤茂一，“インターネットを用いた研究支援システム”，私情協 2003 年発表大会予稿集，pp.72-73, 東京，2003 年 9 月．

[5] 平澤茂一，鴻巣敏之，野村亮，中澤真，松嶋敏泰，“「インターネットを用いた研究活動支援環境」”，経営情報学会 2004 年度秋季全国研究発表大会予稿集，pp.212-215, 名古屋，2004 年 11 月．

[6] Concept Bace Search: http://www.justsystem.co.jp/km/index.html

[7] 株式会社ジャストシステム，ConceptBase カスタマーサーポートサービス スタートアップ CD.

[8] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, New York, Addison-Wesley, 1999.

[9] Rocchio, J. ,“ Relevance Feedback in Information Retrieval, ”*The SMART Retrieval System Experiments in Automatic Docment Processing*, Prentice Hall Inc, (1971).

[10] 酒井哲也，情報検索および情報フィルタリングの高精度化に関する研究，早稲田大学博士論文，2000 年 3 月．