ランダムプロジェクションを用いたニュースストリームの検索

Retrieval for News Stream by Random Projection

大内 浩仁♣　三浦 孝夫♦
塩谷 勇♠

Hirohito OHUCHI　Takao MIURA
Isamu SHIOYA

本論文は効率的なニュースストリームの検索方法を提案する．ニュースストリームの検索においては，更新情報をどのように取り扱うかが検索処理の効率向上において問題となる．ランダムプロジェクションを用いた動的な次元縮小が検索性能および検索時間の双方で充分実用的な処理を可能とすることを述べ，実験によりその有効性を示す．

In this investigation we discuss powerful yet efficient retrieval mechanism for news stream. Difficulty comes from the fact how to manage incremental information while keeping efficiency. Recently *random projection* has been paid much attention on dynamic dimensionality reduction. Here we show this novel technique is really useful for querying news stream in terms of cost and accuracy. We examine some experimental results and excellent efficiency in computation.

1. 前書き

近年, 情報検索の分野において, TDT(Topic Detection and Tracking)プロジェクトに代表される時系列データへの要求が高まっている．特に，文書の時系列データとして配信されるニュース記事，いわゆるニュースストリームから記事を検索するためにはいくつかの問題がある．

一般にニュースストリームには大きく異なる2つの特性が含まれている．新聞のように，新たなトピックが発生すると爆発的な量のニュースが生成され，時間と内容の関連性を追跡することが重視される．これをトピック特性という．これに対し，TVニュース放送では，一定時間に一定量のニュースが定常的に生成される．このことを実時間特性という．ニュースストリームに対する検索では，両特性に関して検索基準や結果の評価が異なる．このため，検索手法の有効性を検証するためには，トピック性および実時間性に対する検討が必要である．また，ニュースストリームに対する検索は即応性を持たなければならない．

一般に，テキスト集合に対する検索は主な単語を取り出し，これらを用いたベクトル空間モデルに基づいて処理される[1]．テキストデータは語いの数だけ次元が存在し，一般的に数万から数十万次元の高次元データとなる．高次元データをそのまま扱うと，計算機容量の確保および即応性への対応が困難になる．計算機容量の効率化と，即応性を実現するためには，テキストデータの次元を縮小して格納する必要がある．

テキストデータにおける次元縮小の方式としてLatent Semantic Indexing(LSI)[2][3][4][5]が知られている．LSI手法では，特異値分解(SVD)を用いて検索空間の次元を縮小する．これにより検索精度を維持したまま次元を大きく縮小することができるため，検索効率と検索精度を両立することができる．しかし，LSI手法にはデータの更新に対してSVDのための再計算が必要となる．LSI手法をニュースストリームの差分情報に対応させるのは容易ではない．

本研究では，ランダムプロジェクション(RP)[6]を用いることによって，計算機容量および検索効率において効率的なニュースストリームの検索方式を提案する．RP手法による次元縮小は計算処理が少なく，データの更新時に再計算を行う必要がないため，差分情報に対して動的に次元縮小を行い，検索質問に対する即応性を保持することができる．

2. テキストデータの次元縮小

テキストデータの次元を縮小する方法として，RP 手法について述べる[6][7][8]．以下ではベクトル空間モデルに基づき，単語数 d，文書数 N の単語・文書行列 X を考える．行列の大きさは d 行 N 列であり，それぞれの列ベクトルが 1 件の文書を表している．行列 X の i 行 j 列の要素 X_{ij} は，文書 j における単語 i の頻度である．

ランダムプロジェクション(RP)行列は要素をランダムに決定した行列である．これにより高次元データを低次元の部分空間に射影する．以下では，大きさ d×N の単語・文書行列 X を大きさ k×N(k << d)の単語・文書行列 X_{RP} に射影する．このため大きさ k×d の RP 行列 R を作成する．単語・文書行列 X の RP 手法による次元縮小は，次の計算で行う．

$$X_{RP}^{k\times N} = R^{k\times d} X^{d\times N} \quad (1)$$

RP 行列 R の要素を構成する際に，非常に単純な要素の分布[6]が提案されている．RP 行列 R の要素 r_{ij} は，次のような独立した分布をとるように並ぶ．

$$r_{ij} = \sqrt{3}\cdot\begin{cases} +1 & (\text{確率}\ \ 1/6) \\ 0 & (\text{確率}\ \ 2/3) \\ -1 & (\text{確率}\ \ 1/6) \end{cases} \quad (2)$$

この分布に基づいて構成される Rでは, 列ベクトルの長さの期待値が全て 1 になる．

質問検索は，質問をベクトル表現した $\mathbf{q}^{d\times 1}$ を k 次元空間に射影して行う．

$$\mathbf{q}_{RP}^{k\times 1} = R\,\mathbf{q}^{d\times 1} \quad (3)$$

検索結果として質問ベクトルと文書ベクトルとの類似度をコサイン尺度で計算し，ランキングを表示する．

RP 手法の次元縮小にともなう誤差は，ベクトル間のユークリッド距離に対して定義される．d×N 行列 Xから任意の 2 つの列ベクトルを取り出し，$\mathbf{x}_1$ および $\mathbf{x}_2$ とおく．

$\mathbf{x}_1$ と $\mathbf{x}_2$ の d 次元におけるユークリッド距離は，$|\mathbf{x}_1 - \mathbf{x}_2|$ で定義される．RP 行列 Rにより k 次元に縮小された空間における $\mathbf{x}_1$ と $\mathbf{x}_2$ のユークリッド距離は以下の式で再現することができる[7].

$$\sqrt{d/k}\,|R\mathbf{x}_1 - R\mathbf{x}_2| \quad (4)$$

式(4)が成り立つためには，Rが直交行列($R^{\mathrm{T}}R=I$)である必

♣ 学生会員　法政大学大学院工学研究科修士課程
i03r3208@k.hosei.ac.jp
♦ 正会員　法政大学工学部情報電気電子工学科
miurat@k.hosei.ac.jp
♠ 正会員　産能大学経営情報学部情報学科
shioya@sanno.ac.jp

要がある．Rの直交性に対する誤差を表すd×d行列εを次の式で定義する．

$$\varepsilon = R^T R - I \quad (5)$$

このときεの要素は，平均0，分散1/kの正規分布をとる[8]．よって縮小次元数kを大きくするほど，誤差は減少する．

LSI手法では，特異値分解(SVD)によって射影行列を求める．単語・文書行列XのSVDは，$X = U \times S \times V^T$で表される．$U$はd×rのユニタリ行列である．$S$はランクrの対角行列でこの対角要素を特異値と呼ぶ．Vはr×Nのユニタリ行列である．次元縮小にはUを射影行列として用いる．SVDは大きい計算量を必要とし，元の行列に依存する処理である．このため更新に対して再計算が必要となる．対してRP行列は少ない計算量で作成可能である，また，データに依存しないため，更新に伴う再計算は必要ない．RP手法は本質的にニュースストリーム検索に対して有効といえる．

3. ニュースストリーム検索の基準

ニュースストリームでは，時系列的に配置されたテキストデータが順次入力されていくことになる．そのため，更新されたばかりの新しい文書と過去の文書が混在する．一般的に，質問者にとっては過去の情報より新しい情報に価値があることが多い．その場合，過去のデータと新しいデータを同列に扱うことはできない．

トピック性ニュースストリームに対する検索は，一般的に最新のトピックに対して行われる．検索要求を満たすためには，新しい文書に対して優先順位を与える仕組みが必要となる．一方，実時間性ニュースストリームでは，過去1週間のニュースなど，検索期間のみ指定する質問が多い．この場合，期間内の文書だけを検索対象とする代わりに，時間差による優先度の格差は存在しない．

本研究では，数学的関数を用いた重み係数を計算することで，ニュースストリームへの検索要求を表現する．トピック性ニュースストリームに対しては，指数関数的に減少する重み係数を使用する．tを文書の経過時間とおくと，重み係数$w_a(t)$は，以下の式によって表される．

$$w_a(t) = \exp(-t/a) \quad (6)$$

aは重み係数の減少度を調節するパラメータである．aが大きいほど重み係数の減少は緩やかになり，a = ∞で減少が完全に止まる．

実時間性ニュースストリームの検索では，窓関数による重み係数を使用して検索期間を限定する．窓関数による重み付けでは，検索期間を表すパラメータをpと置いた場合，重み係数$w_a(t)$を次のような式で表す．

$$w_a(t) = \begin{cases} 1 & (t \le p) \\ 0 & (t > p) \end{cases} \quad (7)$$

いずれの場合も，質問者が関数のパラメータを与えることで，様々な優先順位や検索期間の与え方を制御できる．

4. 実験

本章では，RP 手法を用いたトピック性および実時間性ニュースストリームの検索実験を行う．まず LSI 手法と RP 手法を比較する．次にトピック性および実時間性ニュースストリームの検索を行い，検索結果について考察する．

4.1 実験環境

以下の実験は，FreeBSD4.6.2，Pentium4 2.8GHz，メモリ 1GB の計算機上で行う．ニュースストリームのデータとして，Reuter-21578 コーパス[1]および TDT2 コーパス[2]を使用する．これらのコーパスの詳細を表 1 に示す．

表 1　Reuter コーパスおよび TDT2 コーパスの実験環境
Table.1　Properties of Reuter/TDT2 Corpus

	Reuter コーパス	TDT2 コーパス
特性	記事データ	放送データ(CNN)
有効文書数	19042 件	15785 件
文書期間幅	1 年間	6 ヶ月間
索引語数	2662 語	2859 語
検索間隔	6 時間ごと	1 件ごと
検索回数	199 回	15785 回

Reuter コーパスでは，6 時間ごとに出現する文書数に 0 件から 422 件までの幅がある．この特徴から Reuter コーパスをトピック性ニュースストリームとして扱う．

TDT2 コーパスは実時間性ニュースストリームに対応する．CNN の放送時間が決まっているため，検索期間を区切った場合に検索文書数がほぼ一定になる．

4.2 評価方法

検索結果の評価として，11 点平均適合率を用いる．11 点平均適合率とは，0.0 から 0.1 刻みで 1.0 までの再現率における適合率の平均値である．

再現率は，検索漏れの少なさを示す尺度であり，

$$\frac{\text{検索された文書中の適合文書の数}}{\text{全文書中の適合文書の数}}$$

で表す．適合率は，検索ノイズの少なさを示す尺度であり，

$$\frac{\text{検索された文書中の適合文書の数}}{\text{検索された文書の数}}$$

で表す．再現率と適合率はトレード・オフの関係にあるため，11 点平均適合率が検索精度を示す指標となり得る．

適合文書として，次元縮小を行わない状態で質問検索を行い，その結果類似度が 0.5 以上となった文書を選ぶ．次元縮小による検索精度への影響を調べることができる．

ニュースストリームの検索では，検索質問による 11 点平均適合率の推移を図示し，11 点平均適合率の平均値を求めて総合的な指標とする．

4.3 LSI 手法と RP 手法の比較

Reuter コーパスを用いて，RP 手法と LSI 手法を使用した検索を行う．次元縮小の処理に必要な時間および次元縮小時の検索精度を実験により比較し，考察する．次節以降では，ニュースストリームについての考察のみを行う．

4.3.1 処理時間

LSI 手法による検索は，計算機容量の問題から，先頭の 10000 件のみを処理する．RP 手法による検索では，19042 件の文書を 1 度に処理する．LSI 手法および RP 手法が次元縮小に要した時間を表 2 に示す．

表 2　RP 手法および LSI 手法の処理時間
Table.2　Computation Time in RP/LSI

	次元数	処理時間(秒)
(RP)	100	74
	200	150
	300	231
	400	308
	500	387
(LSI)	2662	21469

処理時間の比較では，RP 手法が圧倒的に勝っている．LSI では，次元数を減らしても処理時間は減少しない．他方，RP

[1] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
[2] http://www.ldc.upenn.edu/Projects/TDT2/

行列は少ない計算量で作成が可能であるとともに，縮小次元数に応じて行列の大きさを減らすことができる．

4.3.2 検索精度

縮小後の次元数を 5 次元から 250 次元までの 10 段階に設定し，それぞれの次元数で LSI 手法と RP 手法による質問検索の検索精度を求める．

RP 手法を用いた検索では，それぞれの次元で 3 回ずつ実験を行い，11 点平均適合率の平均値を最終的な評価とする．同時に RP 手法による次元縮小で生じる分散を計る．

検索の結果を図 1 に示す．

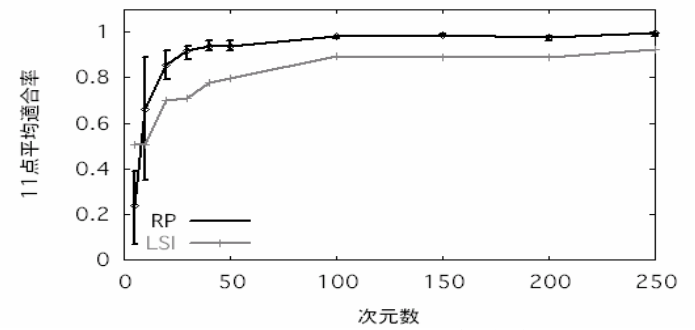

図 1 RP 手法および LSI 手法の検索精度
Fig.1 Accuracy in RP/LSI

検索精度の比較では，平均値のみを考えれば 5 次元以外の全ての次元で RP 手法が上回っている．ベクトル間の距離を保存するという RP 手法の特性が，検索精度の維持に寄与していると考えられる．分散は次元数が増加するほど少なく，100 次元以上では最低値と最大値の差が 1～2%に収束している．よって本実験においては，100 次元以上で RP 手法が安定して LSI 手法と同等以上の検索性能を発揮するといえる．

4.4 RP 手法によるトピック性ニュースストリームの検索

Reuter コーパスに対して指数関数(式(6))に基づく重み付けを行う．t の単位は日数とする．6 時間なら 0.25 である．

過去のデータに対する重み付けのパラメータ a は，急激な重み付け(a=10)，緩やかな重み付け(a=45)，重み付けなし(a=∞)の 3 種類とする．急激な重み付けでは，重み係数は約 7 日間で 0.5 に減少する．緩やかな重み付けでは，約 30 日間で 0.5 に減少する．重み付けなしの場合は，重み係数は常に 1 である．

それぞれの重み付けで 100 次元，300 次元，500 次元の 3 つの次元数における検索質問を行い，11 点平均適合率の推移を求める．急激な重み付けの検索結果を図 2 に，緩やかな重み付けの検索結果を図 3 に，重み付けなしの場合の検索結果を図 4 に示す．

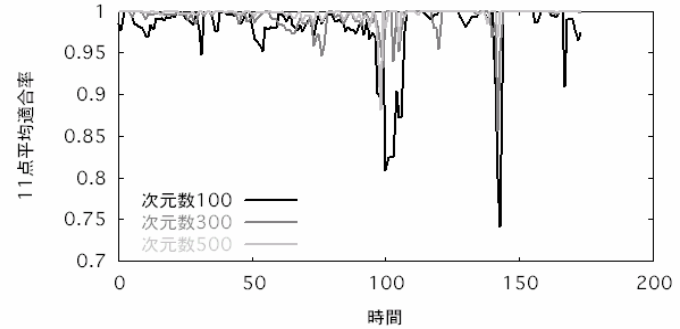

図 2 急激な重み付け(a=10)の検索精度
Fig.2 Query Accuracy with weight exp(t/10)

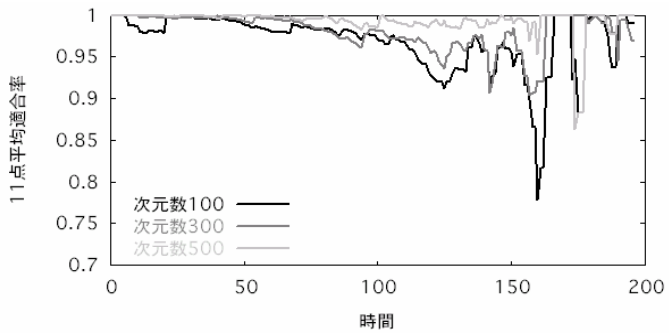

図 3 緩やかな重み付け(a=45)の検索精度
Fig.3 Query Accuracy with weight exp(t/45)

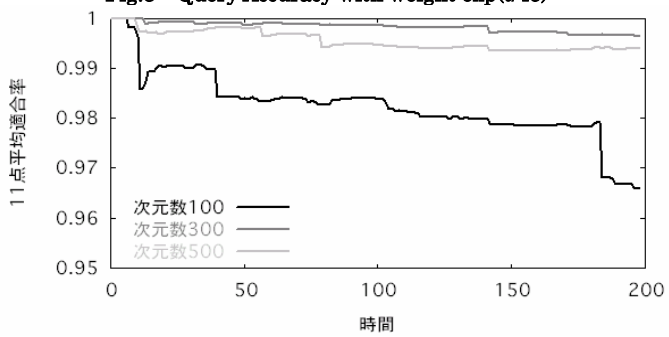

図 4 重み付けなし(a=∞)の検索精度
Fig.4 Query Accuracy with no weighting

トピック性ニュースストリームにおける 11 点平均適合率の平均値を表 3 に示す．

表 3 11 点平均適合率の平均値—Reuter コーパス
Table.3 Average Values of Average Precision of 11 Points in Reuter Corpus

重み付け方法	100 次元	300 次元	500 次元
急激な重み付け	0.968	0.980	0.992
緩やかな重み付け	0.979	0.992	0.997
重み付けなし	0.982	0.998	0.995

4.5 RP 手法による実時間性ニュースストリームの検索

実時間性ニュースストリームとして TDT2 コーパスを用いる．重み付けには窓関数(式(7))を用いる．検索期間のパラメータ p を日数として，1 日間(p=1)，7 日間(p=7)，30 日間(p=30)の 3 種類を与える．前節と同様，100 次元，300 次元，500 次元において検索質問を行い，11 点平均適合率の推移を求める．検索期間 1 日の検索結果を図 5 に，検索期間 7 日の検索結果を図 6 に，検索期間 30 日の場合の検索結果を図 7 に示す．

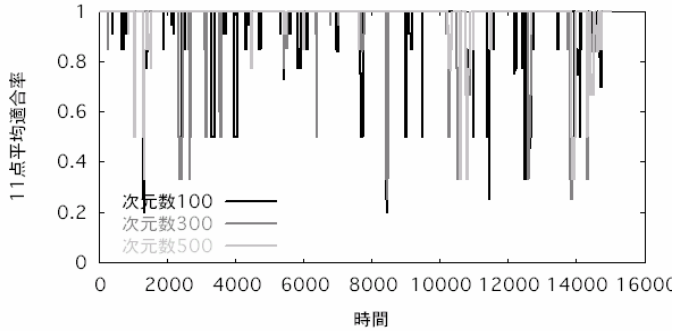

図 5 検索期間 1 日(p=1)の検索精度
Fig.5 Query Accuracy with one day

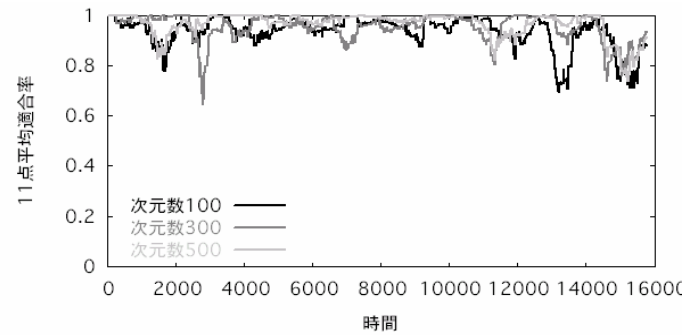

図 6　検索期間 7 日(p=7)の検索精度
Fig.6　Query Accuracy with 7 days

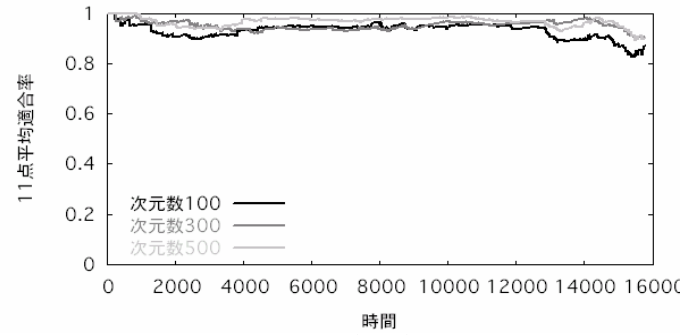

図 7　検索期間 30 日(p=30)の検索精度
Fig.7　Query Accuracy with 30 days

実時間性ニュースストリームの検索に対する 11 点平均適合率の平均値を表 4 に示す.

表 4　11 点平均適合率の平均値—TDT2 コーパス
Table.4　Average Values of Average Precision of 11 Points in TDT2 Corpus

検索期間	100 次元	300 次元	500 次元
1 日	0.957	0.965	0.981
7 日	0.933	0.952	0.965
30 日	0.931	0.951	0.961

5. 考察

トピック性ニュースストリームでは，11 点平均適合率の平均値は全ての場合で 90%を超えており, ニュースの更新に対して安定した検索を行っている．また，次元数が高くなるに従って平均値も上昇する傾向にあり，RP 行列の誤差保証が正しいことを示している.

11 点平均適合率の推移を見ると，重み付けなしの場合では適合率が継続して低下しているのに対し，重み付けがある場合には適合率の低下は局所的になっている．重み付けなしの場合は，検索もれ文書が更新によって除外されることが無いため，適合率が低下し続けると考えられる．検索漏れ文書が残り続ける重み付けなしの場合で最も高い平均値を得たという結果は，全体を通して次元縮小の誤差が小さく抑えられていることを示している.

実時間性ニュースストリームでは，11 点平均適合率の平均値がトピック性と同じく 90%以上の値を保持している. 次元が高いほど平均値が高くなる傾向も同様に見られる．検索期間ごとの比較では，短いほど平均値が高くなっており，トピック性とは逆である.

11 点平均適合率の推移を見ると，検索期間 1 日(図 5)の場合においてごく局所的にかなり適合率の低い検索質問が存在する．この原因として，検索質問ごとの適合文書の数が極端に少ないことが挙げられる．適合文書が 2，3 件しかない場合には，1 件の検索もれが大幅な適合率の減少を引き起こすことになる.

結果的にトピック性, 実時間性ニュースストリームの双方で高い検索精度を得ていることから，RP 手法がニュースストリームの検索にきわめて有効であるといえる.

6. 結び

本研究では，ニュースストリームの検索において RP 手法を適用した．RP 手法における誤差の保証から動的な検索処理が行えることを述べ，これにより検索効率と検索時間を充分実用的な範囲で両立したニュースストリームの検索が可能になることを示した.

今後は使用される記憶域についての議論を加え，モバイル環境下での利用を考えていく予定である.

[謝辞]

本研究の一部は文部科学省科学研究費補助金(課題番号 16500070)の支援をいただいた.

[文献]

[1] 北 研二, 津田 和彦, 獅子堀 正幹: “情報検索アルゴリズム”, 共立出版, 2002.

[2] Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. A.: “Indexing by latent semantic analysis”, journal of the American Society for Information Science, Vol 41, No. 6, pp. 391-407, 1990.

[3] Oh’ Uchi, H. Miura, T. and Shioya, I.: “Retrieval for Text Stream by Random Projection”, International conference on Information Systems Technology and its Applications (ISTA), pp. 151-164, 2004.

[4] Papadimitriou, C. H., Raghavan, P., Tamaki, H. and Vempala, S.: “Latent semantic indexing: A probabilistic analysis”, In Proc. 17th ACM Symp. on the Principles of Database Systems, pp 159-168, 1998.

[5] Berry, M. W., Dumais, S. T. and O'Brien, G, W.: “Using linear algebra for inteligent information retrieval”, SIAM Review, Vol. 37, No. 4, pp. 573-595, 1995.

[6] Achlioptas, D.: “Database-friendly random projections”, In Proc. ACM Symp. on the Principles of Database Systems, pp 274-281, 2001.

[7] Bingham, E. and Mannila, H.: “Random projection in dimentionality reduction: Applications to image and text data”, Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001), pp 245-250, 2001.

[8] Kaski, S.: “Dimensionality reduction by random mapping: Fast Similarity Computation for Clustering”, In Proc. Int. Joint Conf. on Neural Networks (IJCNN), Vol 1, pp. 413-418, 1998.

大内 浩仁　Hirohito OHUCHI
法政大学大学院工学研究科電気工学専攻修士課程在学中．情報検索の研究に従事．日本データベース学会学生会員.

三浦 孝夫　Takao MIURA
京都大学理学部，工学博士(東京大学)．現在，法政大学工学部情報電気電子工学科教授．データモデル，知識表現，演繹データベース，複合オブジェクトなどの分野の研究に従事．電子情報通信学会，ACM 各会員.

塩谷 勇　Isamu SHIOYA
東京電機大学大学院修士課程了．現在，産能大学経営情報学部教授．時系列モデルの同定，論理プログラミング，グラフ文法，論理データベースの研究に従事．電子情報通信学会，ACM 各会員.