情報処理学会論文誌 Vol.56 No.3 1003–1012 (Mar. 2015)

種々のテキスト検索モデルの頑健性向上による音声ドキュメント検索の高精度化

市川 賢[1] 北岡 教英[2,a)] 柘植 覚[3] 武田 一哉[1] 北 研二[2]

受付日 2014年6月30日, 採録日 2014年12月3日

概要：テキスト検索に用いられてきた従来の 3 つの主な検索手法（ベクトル空間モデル，クエリ尤度モデル，適合モデルに基づく手法）に対し統一的な枠組みで改良を加えることで，音声ドキュメント検索における語彙外単語や音声認識誤りに対処する手法を提案し，比較検討を行った．各検索手法に対し，新たな検索質問拡張手法，および音節の 3 連鎖を単語と同様に扱う検索を単語単位の検索とスコアレベルで組み合わせる手法を提案する．提案手法の有効性を NTCIR-9 の SpokenDoc タスクで評価した結果，各手法で Baseline 手法よりも検索性能が向上した．特に，確率に基づくクエリ尤度モデルに基づく手法と適合モデルに基づく手法では検索性能が高かった．提案手法は NTCIR-9 で公表されている公式の最高精度の結果を上回る結果を得た．

キーワード：音声ドキュメント検索，ベクトル空間モデル，クエリ尤度モデル，適合モデル，検索質問拡張，音節認識結果による検索

Improvement of Spoken Document Retrieval based on Various Text Retrieval Models

Ken Ichikawa[1] Norihide Kitaoka[2,a)] Satoru Tsuge[3] Kazuya Takeda[1] Kenji Kita[2]

Received: June 30, 2014, Accepted: December 3, 2014

Abstract: We apply modifications to typical text retrieval methods based on vector space model, query likelihood model, and relevance model, to make them robust to out-of-vocabulary words and misrecognition. We propose novel query expansion methods and combination methods of syllable recognition-based retrieval with word recognition-based retrieval, for these typical methods. We used NTCIR-9 SpokenDoc task to evaluate them. Each modified method achieved better result than baseline. The methods based on stochastic models like the query likelihood model and the relevance model achieved better performance than the vector space model. The performance of our proposed methods was better than the best result published in NTCIR-9 competition.

Keywords: spoken document retrieval, vector space model, query likelihood model, relevance model, query expansion, syllable recognition-based retrieval

1. はじめに

インターネットの普及により，容易に大量のテキスト文書が入手可能となり，それらを効率的に検索するためのテキスト文書検索の研究がさかんに行われ，多くの検索エンジンが実用化に至っている．近年では，音声 · 画像 · ビデオの記録 · 編集機器の高性能化にともない，YouTube などの動画サイトや Podcast が普及し，音声を含むマルチメディアデータが我々の身近な存在となり，それらに対する検索要求が高まっている．こうしたデータの場合には，データ（音声ドキュメント）中の音声に対し大語彙連続音声認識を行い，音声をテキストデータとして書き起こし，それらに対し，テキストで検索質問を与え検索する方法が可能であ

[1] 名古屋大学
Nagoya Uniersity, Nagoya, Aichi 464–8603, Japan
[2] 徳島大学
The Uniersity of Tokushima, Tokushima 770–8506, Japan
[3] 大同大学
Daido Uniersity, Nagoya, Aichi 457–8530, Japan
[a)] kitaoka@is.tokushima-u.jp

る．このような音声言語情報を対象とした検索技術は「音声ドキュメント検索」と呼ばれ，必要不可欠な技術となりつつある [1]．

しかし，従来の検索手法では，検索対象のテキストに誤りが含まれていることを仮定していないので，音声認識で生じる認識誤りを扱うことができない．特に，音声認識における未知語は，音声認識結果の書き起こし文書に現れないため，検索することができない．音声ドキュメントを対象とする検索では，音声認識時に生じるこれらの現象を，検索手法でどのように扱うかが課題となる．

これまでに音声ドキュメント検索に関する，数多くの検索手法が提案されてきた．たとえば，連続音節認識を用いることによるサブワード単位でのインデックスを用いる研究があげられる [2]．また，岩見ら [3] はあらかじめ誤りを考慮した音節 n-gram インデックスを構築することによる高速な音声ドキュメント検索手法を提案している．認識誤りの対処には，音声認識結果のラティスやコンフュージョンネットワークから構築したインデックスを用いた検索を行う研究もある [4]．これらの研究ではいかに単語単位の検索とサブワード単位の検索を組み合わせて頑健な検索を実現するかが鍵となる．本研究では種々の基本的な単語単位の検索との組合せを実現してその効果を示す．

さらに，Web 上の情報を用いることによって，検索質問や検索対象となる音声ドキュメントの拡張を行う研究がされている．Terao ら [5] は，検索に用いる音声検索質問に関連性のある Web ページの情報を利用することによって，検索質問拡張を行っている．杉本ら [6] は検索対象となる音声ドキュメントと類似性の高い Web ページを用い，音声認識結果から作成されたインデックスと Web ページから作成されたインデックスを適宜組み合わせて利用することにより性能の向上を図っている．増村ら [7] は確率的言語モデルに着目し，Web 関連文書を用いた文書モデル拡張手法を提案している．また確率的言語モデルを用いた音声ドキュメントの検索については Chen [8] がトピックモデル [9] に着目し，PLSA [10] との比較を通じ WTM（Word Topic Model）という単語レベルでのトピックモデルを提案している．以上のように，音声ドキュメント検索では，いかに認識誤りや未知語に対処するかが重要である．これらの研究が示すとおり，Web ページを利用した検索質問拡張は効果的である．これまでに DP マッチングベースの方法 [5] やクエリ尤度モデルベースの方法 [7] など，いろいろな方法に個々に導入されてきた．本論文では，統一的な方法で種々の基本的な手法と組み合わせて効果を比較する．

ここまでに述べてきたように，本論文では従来の情報検索の基本的なモデル（ベクトル空間モデル，クエリ尤度モデル，適合モデル）に基づく手法に対し統一的な改良を加え，未知語や認識誤りに対処する．まず未知語に対し，新たな検索質問拡張手法を用いる．検索質問拡張として，検索質問と，検索質問による Web 検索結果から，検索対象をより精細にモデル化する手法を提案する．また未知語と認識誤りに対し，音節単位の認識に基づく検索を用いる．音節認識結果の音節 3 連鎖を索引語として用い，ランキングの際のスコアとして単語認識結果のスコアと効果的に組み合わせる手法を提案する．

2. テキスト検索における主な検索手法

テキスト検索では文書をランキングする方法として代表的な検索手法がある．本章では，本論文で用いる 3 つの基本的な検索手法について述べる．

2.1 ベクトル空間モデルに基づく手法

ベクトル空間モデルに基づく手法 [11] は，検索対象となる文書および検索質問を，索引語となる単語を軸として構成されるベクトルで表現し，その文書ベクトルと検索質問ベクトルの距離に基づきランキングを行う手法である．

検索対象となる文書と検索質問を形態素解析し，任意の品詞の語を索引語として用いる．これらの索引語をベクトルの要素として用い，単語の出現頻度を考慮した以下の重み付けをし，文書と検索質問をベクトルで表現する．

- TF（Term Frequency）
  TF w_T は式 (1) で表される．

$$w_T(t_{ij}) = \frac{n(t_{ij})}{\sum_i n(t_{ij})} \tag{1}$$

  ここで，t_{ij} は文書 j の i 番目の単語を表し，$n(t_{ij})$ は t_{ij} の文書中での生起回数を表す．
- TF-IDF
  TF-IDF は TF に単語の逆文書頻度（Inverse Document Frequency; IDF）を乗じた重みである．検索対象の全文書数を D，単語 t_i が含まれる文書数を $m(t_i)$ とすると，単語 t_i の IDF 値 w_I は式 (2) のように計算される．

$$w_I(t_i) = \log\left(\frac{D}{m(t_i)}\right) \tag{2}$$

  TF-IDF は単語の TF 値に IDF を乗じた重みであるので，式 (3) で計算される．

$$d_{ij} = w_T(t_{ij}) \times w_I(t_i) \tag{3}$$

  ここで d_{ij} は文書 j における i 番目の単語の TF-IDF 値を表す．
- 2 進値
  2 進値は文書中に単語が存在する場合に 1，存在しない場合に 0 となる．

$$d_{ij} = \begin{cases} 1, & \text{if } w_T(t_{ij}) > 0 \\ 0, & \text{if } w_T(t_{ij}) = 0 \end{cases} \tag{4}$$

検索質問ベクトルと文書ベクトルの距離の近さに基づい

て文書のランキングを行う．ベクトル間の距離の計算には以下のコサイン距離を用いる．

$$\begin{aligned} S(\mathbf{d}_j, \mathbf{q}) &= 1 - \cos(\mathbf{d}_j, \mathbf{q}) \\ &= 1 - \frac{\mathbf{d}_j^{\mathrm{T}}\mathbf{q}}{\sqrt{\mathbf{d}_j^{\mathrm{T}}\mathbf{d}_j}\sqrt{\mathbf{q}^{\mathrm{T}}\mathbf{q}}} \end{aligned} \tag{5}$$

ここで S はコサイン距離を表し，$\mathbf{d}_j$ は j 番目の文書の文書ベクトル，$\mathbf{q}$ は検索質問ベクトルを表す．

2.2 クエリ尤度モデルに基づく手法

本節では検索対象文書が検索質問に適合する確率に基づく手法について述べる．ここでは，文書 d が検索質問 q に適合する確率 $P(d|q)$ を求め，この確率の大きさに従い文書のランキングを行う．ベイズの定理より

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \propto P(q|d) \tag{6}$$

となる．ここで，$P(q)$ は文書に依存しないのでランキングに影響せず，定数と見なし，また $P(d)$（文書が適合である事前確率）は，すべての文書に対し一様な値とした．

$P(q|d)$ を推定するにあたり，文書 d を直接取り扱うことはできないので，文書 d の単語は独立に出現すると仮定し，ユニグラム言語モデル θ_d で近似する．そして，文書 d のユニグラム言語モデル θ_d から検索質問 $q = q_1, \ldots, q_{|q|}$（$|q|$ は質問 q 中の単語数）が生成される尤度 $P(q|\theta_d)$ を式 (7) で求める．

$$P(q|\theta_d) = \prod_{l=1}^{|q|} P(q_l|\theta_d) \tag{7}$$

この $P(q|\theta_d)$ はクエリ尤度と呼ばれ，検索対象文書から検索質問が生成される確率を表している．

式 (7) は，検索質問 q において単語 $w_i \in V = \{w_1, \ldots, w_{|V|}\}$ が出現する回数 $c(w_i, q)$ を用いて次のように書ける．

$$P(q|\theta_d) = \prod_{w_i \in V} P(w_i|\theta_d)^{c(w_i,q)} \tag{8}$$

このクエリ尤度を算出し，その大きさに基づきランキングを行う．

ユニグラム言語モデル θ_d は，ごく一部の単語に対してのみ値を持つスパースなモデルである．この零頻度問題に対処するために，本論文ではディリクレスムージング [12] を用いた．

$$\begin{aligned} P(w_k|\hat{\theta}_d, \mu) &= \frac{c(w_k, d) + \mu P(w_k|\theta_C)}{|d| + \sum P(w_k|\theta_C)} \\ &= \frac{c(w_k, d) + \mu P(w_k|\theta_C)}{|d| + \mu} \\ &= \frac{|d|}{|d| + \mu}\frac{c(w_k, d)}{|d|} + \frac{\mu}{|d| + \mu}P(w_k|\theta_C) \\ &= \frac{|d|}{|d| + \mu}P(w_k|\theta_d) + \frac{\mu}{|d| + \mu}P(w_k|\theta_C) \end{aligned} \tag{9}$$

ここで，θ_C はコーパス全体から求めたユニグラム言語モデル（コレクションモデル），$|d|$ は文書 d の単語数であり，$\hat{\theta}_d$ はディリクレスムージングによって得られた係数 μ を用いて得られた新たな言語モデルである．

2.3 適合モデルに基づく手法

ある検索質問に対し，適合であるクラス（$R = 1$）に属する文書群を用いて推定された言語モデルを，適合モデル（Relevance model）[15] と呼ぶ．ここでは，以下の仮定をする．

- ユーザが検索したいと考えている要求には，その根底に潜在する適合モデル R が存在する．
- 検索質問単語は適合モデルから生成されると考え，適合モデルから検索質問単語が生成される確率を $P(q|R)$ とする．
- 適合文書も同様に適合モデルから生成されると考え，適合モデルから文書が生成される確率を $P(d|R)$ とする．

さらに，ここで文書や検索質問に存在する語 w が適合モデルから独立に生成されると仮定すると，$P(w|R)$ を推定する問題に帰着する．検索質問 $q = q_1, \ldots, q_k$ と適合文書に存在する単語が，R から生成される過程を，未知のブラックボックス R からランダムに単語をサンプリングすると考える．k 回サンプリングをしたら，単語列 $q_1, \ldots, q_k$ を観測したとし，次の試行で単語 w が観測される確率として妥当と考えられるのは単語列 $q_1, \ldots, q_k$ を観測した元で w を観測する条件付き確率

$$P(w|R) \approx P(w|q_1, \ldots, q_k) \tag{10}$$

である．確率の定義から条件付き確率を，単語 w と検索質問 $q_1, \ldots, q_k$ を観測する同時確率で表現できる．

$$P(w|R) \approx P(w|q_1, \ldots, q_k) = \frac{P(w, q_1, \ldots, q_k)}{P(q_1, \ldots, q_k)} \tag{11}$$

よって，推定すべき確率 $P(w|R)$ は，単語 w と検索質問単語列 $q_1, \ldots, q_k$ を観測する同時確率の推定問題に帰着する．

一般に，検索質問 $q_1, \ldots, q_k$ を用いて Web 検索などにより得られた上位 J 文書のユニグラムの集合 $M_j \in \mathcal{M}$, $(j = 1, \ldots, J)$ を用いて，式 (12) で表す．

$$\begin{aligned} &P(w, q_1, \ldots, q_k) \\ &\quad = \sum_{M_j \in \mathcal{M}} P(M_j)P(w, q_1, \ldots, q_k|M_j) \\ &\quad = \sum_{M_j \in \mathcal{M}} P(M_j)P(w|M_j)\prod_{i=1}^{k} P(q_i|M_j) \end{aligned} \tag{12}$$

確率の加法性を満たすために

$$P(q_1,\ldots,q_k)=\sum_w P(w,q_1,\ldots,q_k) \tag{13}$$

とする．これらは，零頻度問題の回避のために以下の線形補間によるスムージング

$$P(w|M_j,\rho)=\rho P(w|M_j)+(1-\rho)P(w|\theta_C) \tag{14}$$

（ここで，$P(w|\theta_C)$ はコーパス全体のユニグラム）を行った上で計算される．

文書のランキングは，推定した単語の生成確率分布 $P(w|R)$ と，検索対象文書のユニグラムモデル $P(w|\theta_d)$ の全単語に対する分布間距離を，以下のように KL-ダイバージェンスで測り，その負の大きさによりランキングを行う [16]．

$$\begin{aligned}-KL(R||\theta_d)&=\sum_{w_i\in V}P(w_i|R)\log P(w_i|\theta_d)\\&\quad-\sum_{w_i\in V}P(w_i|R)\log P(w_i|R)\\&\propto\sum_{w_i\in V}P(w_i|R)\log P(w_i|\theta_d)\end{aligned} \tag{15}$$

3. 検索質問拡張による改良

本節では検索質問拡張手法について述べる．検索質問拡張とは，シソーラスなどの単語間の関係を用いたり，検索エンジンを用いて Web から収集した Web ページに含まれる単語を用いるなどにより，検索質問に含まれない単語を追加する手法である．検索質問拡張を行うことにより，検索質問に現れていないが，検索質問内の語と共起しやすい単語を用いた検索が可能となり，検索性能の向上が期待できる．

3.1 ベクトル空間モデルに基づく手法

検索質問拡張は，検索質問の単語ベクトルを，Web などから得た文書を用いて拡張する方法である．Web から取得してきた文書に対し形態素解析を行い，検索質問と同様にしてベクトルを作成する．元の検索質問から作成したベクトルを $\mathbf{q}_o$，拡張用の文書から作成した拡張ベクトルを $\mathbf{q}_e$ とする．検索質問拡張手法 [13] は，$\mathbf{q}_o$ と $\mathbf{q}_e$ を用い以下のように拡張検索質問ベクトルを作成する．

$$\hat{\mathbf{q}}=(1-\alpha)\mathbf{q}_o+\alpha\mathbf{q}_e \tag{16}$$

ここで $\hat{\mathbf{q}}$ は拡張検索質問ベクトルを表し，α はベクトル間の線形補間係数であり，どちらのベクトルに重みを置くかを表す．α は経験的に決められる[*1]．

[*1] 我々は，α を経験的に定める必要のない，$\mathbf{q}_o$ と $\mathbf{q}_e$ で張られる平面による検索対象文書モデリング手法を提案している [17]．しかし，本論文中の実験では従来法に性能が及ばなかったために比較実験からは割愛した．

3.2 クエリ尤度モデルに基づく手法

Ponte [14] が提案した検索質問拡張手法は，元の検索質問を用い，Web から取得してきた上位 M 文書に含まれる N 単語を，元の検索質問に検索語として加えるという手法である（先行研究では $M=5$，$N=5$）．追加する単語は式 (17) で示すスコア L_w が大きな値となる上位 N 単語を追加する．

$$L_w=\sum_d\log\frac{P(w|M_d)}{P(w|\theta_C)} \tag{17}$$

ここで，M_d は Web から取得した文書 d によるユニグラム言語モデルを表し，$\sum_d$ は Web から得られた文書すべてに対して加算することを意味する．また，θ_C はコーパスのユニグラム言語モデルを表す．L_w が大きな値となる単語は，検索質問に関連度が高い単語であると考えられる．この手法は，追加する単語数や，文書数により性能が左右される．本論文では，Web から取得する全単語を用い，もとの検索質問単語と効率的に組み合わせることで検索性能の向上を図る．

元の検索質問を q_0，拡張用の単語を q_e とし，拡張検索質問を $\hat{q}$ とすると，式 (8) は式 (18) となる．

$$P(\hat{q}|d)=\prod_{w_i\in V}P(w_i|\theta_d,\mu)^{c(w_i,q_o)}\prod_{w_i\in V}P(w_i|\theta_d,\mu)^{c(w_i,q_e)} \tag{18}$$

一般に，拡張用単語の単語出現頻度 $c(w_i,q_e)$ は検索質問単語の単語出現頻度 $c(w_i,q_o)$ が 1〜10 のオーダーであるのに対し，数十〜数百と非常に大きな値となる．そこで以下のように，検索質問単語，拡張用単語の出現頻度を，それぞれの単語数で割ることで正規化する．

$$P(\hat{q}|d)=\prod_{w_i\in V}P(w_i|\theta_d,\mu)^{\frac{c(w_i,q_o)}{C_o}}\prod_{w_i\in V}P(w_i|\theta_d,\mu)^{\frac{c(w_i,q_e)}{C_e}} \tag{19}$$

ここで $C_o=\sum_{w_i\in V}c(w_i,q_o)$，$C_e=\sum_{w_i\in V}c(w_i,q_e)$ である．さらに，検索質問か，拡張用単語かどちらを重視するかを表現する補間係数 λ を用い

$$\begin{aligned}&P(\hat{q}|d)\\&=\prod_{w_i\in V}P(w_i|\theta_d,\mu)^{(1-\lambda)\frac{c(w_i,q_o)}{C_o}}\prod_{w_i\in V}P(w_i|\theta_d,\mu)^{\lambda\frac{c(w_i,q_e)}{C_e}}\\&=\prod_{w_i\in V}P(w_i|\theta_d,\mu)^{(1-\lambda)\frac{c(w_i,q_o)}{C_o}+\lambda\frac{c(w_i,q_e)}{C_e}}\\&=\prod_{w_i\in V}P(w_i|\theta_d,\mu)^{\left(\frac{c(w_i,q_e)}{C_e}-\frac{c(w_i,q_o)}{C_o}\right)\lambda+\frac{c(w_i,q_o)}{C_o}}\end{aligned} \tag{20}$$

とする．式 (20) に従い検索対象文書が拡張検索質問に適合である確率を求める．

3.3 適合モデルに基づく手法

適合モデルに基づく手法は，適合モデル R を推定する

際に Web 文書に含まれる全単語を用いて推定しているので，大規模な検索質問拡張を行っていると考えることができる．

この手法は，Web から取得してきた適合文書の振舞いはとらえているが，元の検索質問の情報を捨ててしまっている．適合モデルの推定に，検索質問の情報を取り込むことができればよりよい推定が行える可能性がある [18]．適合モデルの推定において，元の検索質問の情報を保持し利用するために，元の適合モデルと検索質問のユニグラム言語モデル（以下，検索質問モデル）を式 (21) で線形補間を用い組み合わせる．

$$P_l(w|R) = \phi P(w|Q) + (1-\phi)P_o(w|R) \tag{21}$$

ここで，$P_o(w|R)$ は 2.3 節で求めた元の適合モデルを表し，$P(w|Q)$ は検索質問モデルを表し最尤推定を用いて推定する．また $P_l(w|R)$ は元の適合モデルと検索質問モデルを組み合わせた適合モデルを表す．ϕ は線形補間係数を表し，経験的に決められる．分布間距離で文書をランキングする際に，元の検索質問単語に対する $P_l(w|R)$ の値が大きくなるので，元の検索質問単語の影響が大きくなると考えられる．

しかし検索質問は数単語しか含まれていないので検索質問モデルの分布は非常に偏っていると考えられる．そこで，元の検索質問に以下のスコアが大きくなる単語を追加し，関連単語を用い検索質問モデルを拡張する手法を提案する．

$$L_w = \sum_d \log \frac{P(w|M_d)}{P(w|\theta_C)} \tag{22}$$

ここで M_d は Web から取得した文書 d のユニグラム言語モデルを表し，θ_C はコーパスのユニグラム言語モデルを表す．この値が大きいものほど文書に関連した単語であり，検索質問の単語だけでなく，関連単語も考慮することで検索性能の向上が期待できる．

4. 音節認識結果に基づく検索の併用

本章では，音節認識結果に基づいて，音節の 3 連鎖を単語のように扱い検索する手法を，単語単位での検索手法と併用する方法について述べる．なお，本論文では，前章で説明した検索質問拡張を行った場合の単語単位の検索を前提とし，それに本章の方法を併用することとしているが，必ずしも我々の単語単位の方法に限定された方法ではないことを申し添えておく．

4.1 ベクトル空間モデルに基づく手法

4.1.1 距離の線形補間による異なるベクトル空間の併用

音声ドキュメントや検索質問をベクトルで表現する際，索引語として単語，音節など，索引語重みとして TF-IDF や二値重みなどを用いることができ，各々でベクトル空間を構成することが可能である．そのため，各ベクトル空間で，文書ベクトルと検索質問ベクトルの距離を計算することが可能である．そこで，音声ドキュメントを複数のベクトル空間で表現し，各空間で距離を計算し，それらの距離を統合する手法を提案する．音声ドキュメントを複数の異なった空間で表現することは，文書に含まれる情報を多面的に表現することと考えられ，これらの情報を効果的に統合可能であれば，高性能な検索が期待できる．

本論文では，各ベクトル空間で求められる距離を，式 (23) で示す線形補間で統合する．

$$S_{\{\mathbf{q},\mathbf{d}_j\}} = \sum_k \beta_k s_k(\mathbf{q}, \mathbf{d}_j) \tag{23}$$

ここで，$S_{\{\mathbf{q},\mathbf{d}_j\}}$ は距離を統合した際の文書ベクトルと検索質問ベクトル間の距離を表す．また，$s_k(\mathbf{q}, \mathbf{d}_j)$ はベクトル空間 k で計算される文書ベクトル $\mathbf{d}_j$ と検索質問ベクトル $\mathbf{q}$ の距離を表し，β_k は距離統合を行う線形補間係数を表す．従来，索引語検出手法として連続音素認識結果と連続単語音声認識結果を組み合わせる手法として文献 [20] などが提案されているが，音声ドキュメント検索において距離を統合する検討は行われていない．そこで，本論文では式 (23) に示す方法で距離を統合することを提案し，有効性を検証する．

4.1.2 音節認識結果に対する潜在意味解析の適用

潜在意味解析（LSA：Latent Semantic Analysis）[19] は，言葉の同義性などに対処するために発展した統計的技法である．潜在意味解析を行うことで，類似の意味がさまざまな言葉で表現されるという性質を，行列の分解によって取り除き，語句の背後に共通して存在する意味的構造を抽出することができると考えられる．

一般的に潜在意味解析は言葉の同義性に対処するために，単語を索引語として構成された単語–文書行列に対して適用される．本論文では，この潜在意味解析を，音節 3 連鎖を索引語とした音節–文書行列に対し適用することを提案する．音節はもともと意味を持たないひらがな列であるが，単語の部分音節列単位で共起関係を学習することで，同じ単語内や共起しやすい単語内の音節列が近い位置に射影され，単語の意味を利用して空間上に配置されることが期待される．つまり，検索質問に現れないが同義や類義で検索の対象となる語が，未知語や音声認識誤りによって単語として正しく文書中に現れない場合にも，その部分系列が音節認識の結果に現れていれば，検索質問の音節部分列との「意味的」類似性を考慮して検索を行える可能性がある．

4.2 クエリ尤度モデルに基づく手法

索引語として，単語と同様に音節 3 連鎖を用い検索対象文書が検索質問に適合する確率を求めることが可能である．索引語が単語，音節 3 連鎖であった場合の検索対象文書が検索質問に適合する確率をそれぞれ $P_w(d|q)$，$P_s(d|q)$

とすると式 (24) でスコアを組み合わせる．

$$L = (1-\gamma)\log P_w(d|q) + \gamma \log P_s(d|q) \quad (24)$$

ここで，L はランキングに用いる最終的なスコアを表し，γ は線形補間係数を表す．

4.3 適合モデルに基づく手法

適合モデルに基づく検索手法を用いる場合，Web から取得した文書から適合モデルを推定する必要がある．そこで，Web から取得した文書に対し，形態素解析を行い，出力されるひらがな列を用い，そのひらがな 3 つ組を音節 3 連鎖の索引語として用いる．索引語に単語，音節 3 連鎖を用いた場合の分布間距離をそれぞれ $KL_w(R||\theta_d)$，$KL_s(R||\theta_d)$ とするとランキングに用いるスコアを式 (25) とする．

$$r = (1-\nu)KL_w(R||\theta_d) + \nu KL_s(R||\theta_d) \quad (25)$$

ここで，r は最終的なランキングに用いるスコアを表し，ν は線形補間係数を表す．

5. 評価実験

5.1 実験条件

音声ドキュメント検索の検索対象文書として，NTCIR（NII-Test Collection for IR）-9 の音声ドキュメント検索タスク（SpokenDoc）の文書検索（Spoken Document Retrieval; SDR）サブタスクを用いた．

SpokenDoc タスクでは，前段の音声認識と後段の検索性能を区別して評価するために，参加者共通で利用できる音声認識結果が提供されている．これにより，誤りを含むテキストに対する検索手法に焦点を絞った参加が可能となる．提供される認識結果は，単語ベース音声認識による書き起こしと，音節ベース音声認識による書き起こしの 2 通りがある．単語ベース音声認識による書き起こしは，**表 1** の認識条件の下で作成されている．単語認識結果には形態素解析が施され，任意の品詞を索引語として用いることが可能であるが，本論文では，索引語として名詞，アルファベット，カタカナ語を用いた．音節ベース音声認識による書き起こしは，日本語の全音節を辞書として，音節 3-gram を用いて作成されている．書き起こしをオープンな条件の下で作成するために，音響 · 言語モデルの学習データを ID の偶奇にわけて行い，各文の認識時には偶奇の異なるモデルを用いている．単語認識精度，音節認識精度はそれぞれ 70%，75%である．

表 1 音声認識条件
Table 1 Conditions for speech recognition.

音声認識器	Julius 4.1.3
音響モデル	triphone （32 混合，3,000 状態）
使用特徴量	38 次元 （MFCC12 次元+Δ12 次元 +ΔΔ12 次元+Δpower +ΔΔpower，CMN あり）
言語モデル	単語 3-gram
形態素解析ソフトウェア	chasen 2.4.4
形態素解析辞書	UniDic-1.3.9
単語辞書語彙サイズ	約 27,000 単語

SpokenDoc の検索対象文書は，CSJ（Corpus of Spontaneous Japanese）に含まれる音声からなる．CSJ に含まれるデータのうち，学会講演および模擬講演を検索対象文書とし，両講演データを合わせると 2,702 講演となる．

検索質問拡張に用いる文書の取得は，検索質問に含まれる名詞を用いインターネット上の Web を検索し収集した HTML 文書から作成した．Web 検索には API 経由で Google 検索 [21] を用いた．

連続音節認識結果を用いた場合，NTCIR-9 オーガナイザにより提供された連続音節認識結果に対し．小文字（“ゃ”，“ゅ”，“ょ” など）は大文字（“や”，“ゆ”，“よ” など）に変換し，また長音を示す “ー” は除去した．そして変換を行った音声ドキュメントから連続 3 文字（音節 3 連鎖）を抽出し，索引語として用いた．異なり 3 連鎖数は 169,363 であった．

SpokenDoc タスクでは開発セット（以下 Dry Run）として 39 の検索質問文が与えられ，評価セット（以下 Formal Run）として 86 の検索質問文が与えられる．音節認識結果を用いる場合，検索質問を手動でひらがな列に書きおこし，音節認識結果と同様に小文字の大文字変換や長音の除去を行い，音節 3 連鎖を索引語として用いた．

検索性能の評価は式 (26) で示す適合率の平均（Mean Average Precision; MAP）を用いた．

$$M = \frac{1}{Q}\sum_{q=1}^{Q}\overline{P_q} \quad (26)$$

ここで，Q は検索質問文数を示し，$\overline{P_q}$ は検索質問 q における平均適合率を示す．平均適合率は式 (27) で計算する．

$$\overline{P_q} = \frac{1}{R_q}\sum_{k=1}^{R_q}\frac{O(r_k)}{k} \quad (27)$$

ここで，R_q は検索質問 q における適合文書数を示す．また，$O(r_k)$ は，適合文書 r_k が検索された順位を示す．ただし，適合文書は検索順位で昇順に並べ替えられている（$C(r_1) < C(r_2) < \ldots < C(r_{R_q})$，C(∗) は検索順位を表す）とする．本論文では，NTCIR の規定に従い，検索結果上位 1,000 件を用い評価するため，上位 1,000 件以内に含まれない適合文書は式 (27) の計算に使用しない．

以降，基本的には評価には Dry run を用いてる．最後に 5.4 節において手法間の比較を行う際には Formal run の結果も用いる．

5.2 検索質問拡張の効果

5.2.1 ベクトル空間モデル

3.1 節で述べた検索質問拡張手法を用いて検索実験を行った．検索質問拡張は予備実験から検索性能の高かった単語 TF–IDF に対して行った．**図 1** に，Web から検索によって取得した上位文書の数を最も性能の高い時の文書数（予備実験より 26 とした）に固定し，α の値を $0 \leq \alpha \leq 1$ で 0.1 刻みに変化させたときの検索性能を示す．検索質問拡張によって検索性能の向上が見られ，$\alpha = 0.9$ で最高となった．

5.2.2 クエリ尤度モデル

図 2 に，Web から検索によって取得した上位文書の数を，予備実験で適切とされた文書数（ここでは 7）に設定し，式 (20) の λ を $0 \leq \lambda \leq 1$ で 0.01 刻みに変化させたときの提案手法の結果を示す．図から，λ の値を変化させると検索性能が向上していることが分かる．$\lambda = 0.76$ のとき最も性能が高く MAP は 0.455 となった．

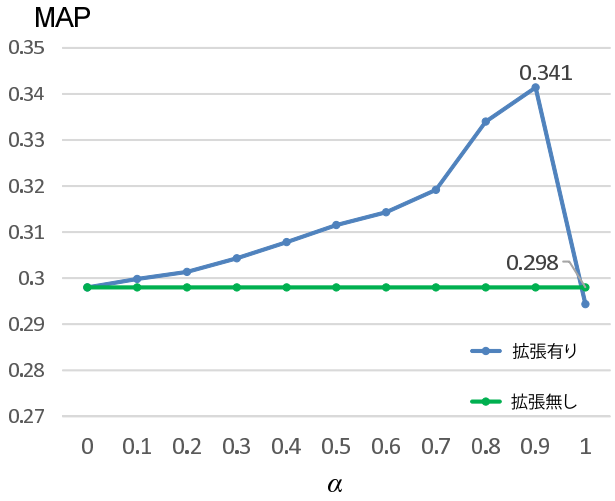


図 1 ベクトル空間モデルにおける α の値による検索性能の変化（Dry Run）

Fig. 1 Retrieval performance of vector space mdoel-based method for various α (Dry Run).

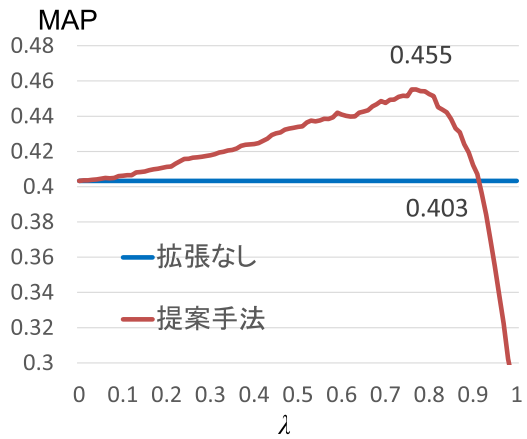


図 2 クエリ尤度モデルにおける検索質問拡張（提案法）の結果（Dry Run）

Fig. 2 Retrieval performance of query likelihood model-based method with query expansion for various α (Dry Run).

5.2.3 適合モデル

適合モデルに提案する検索質問拡張手法を適用した検索実験を行った．予備実験から，適合モデルの推定に用いる Web 文書数は 50 文書とした．

提案手法では式 (21) の元の適合モデルと検索質問モデルとの線形補間係数 ϕ および ρ を調整する必要がある．ρ については，予備実験の結果から $\rho = 0.5$ に決定した．その上での実験結果として，**表 2** に従来手法，提案手法の最も性能が高かったときの各パラメータと検索性能を示す．提案手法を用いることで，従来手法を上回る結果が得られることが分かる．適合モデルを推定する際に，関連単語の生成確率が大きくなることで，検索性能が向上することが分かった．

5.3 音節認識結果に基づく検索の併用の効果

本節では，前節の方法を，さらに音節認識結果に基づく音節 3 連鎖単位の検索と統合することの効果を評価する．

5.3.1 ベクトル空間モデル

図 3 に音節 3 連鎖の TF-IDF で作成した音節–文書行列に対し潜在意味解析を適用し，次元数を変化させたときの結果を示す．図から潜在意味解析を行い，次元を削減すると検索性能が向上し，768 次元のときに最も性能が高くな

表 2 適合モデルに基づく手法の結果（Dry Run）

Table 2 Retrieval performance of relevance model-based method (Dry Run).

手法	ϕ	MAP
検索質問モデルの線形補間	0.1	0.445
検索質問拡張（提案手法）	0.2	0.474
Baseline	0.0	0.414

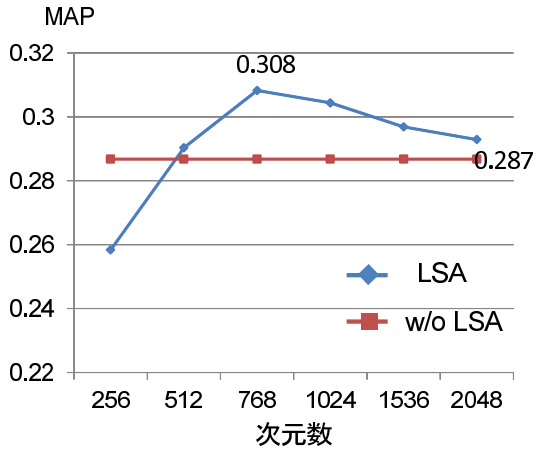


図 3 ベクトル空間モデルにおける音節単位の潜在意味空間での検索性能（Dry Run）

Fig. 3 Retrival performance of vector space model-based method in latent semantic space (Dry Run).

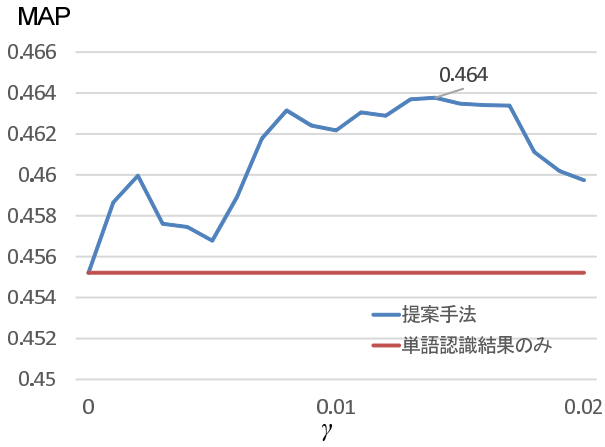


図 4 クエリ尤度モデルにおける音節認識結果の併用（Dry Run）
Fig. 4 Effect of use of syllable recognition result in query likelihood model-based method (Dry Run).

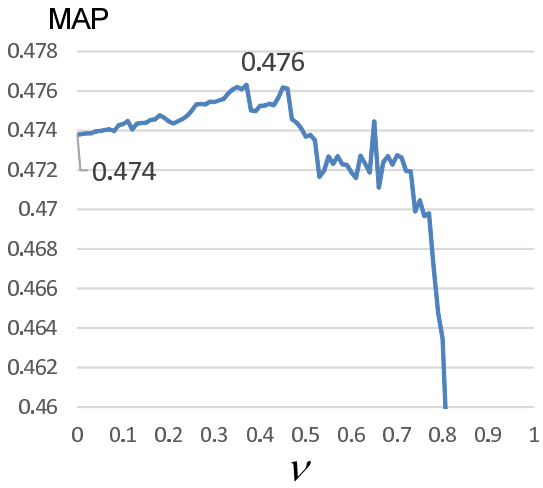


図 5 適合モデルにおける音節認識結果の併用（Dry Run）
Fig. 5 Effect of use of syllable recognition result in relevance model-based method (Dry Run).

ることが分かる．このことから，音節に対し潜在意味解析を適用することが有効であることが分かる．潜在意味解析を音節 3 連鎖に適用することで，もともと意味を持たない音節 3 連鎖が単語の断片として扱われ，単語の意味を利用して潜在意味空間上に配置される効果があると考えられる．

4.1.1 項の方法により，TF-IDF（クエリ拡張あり）の空間，単語二値の空間，音節 3 連鎖の TF-IDF の潜在意味空間における距離を統合した．統合の重み β_k については，さまざまな値の組み合わせを Dry run で試した結果，最も良い結果となったものとした．その場合の β_k は，単語 TF-IDF，単語二値，音節 3 連鎖それぞれ 0.2，0.7，0.1 であった．このときの MAP 値は 0.394 となった．

5.3.2 クエリ尤度モデル

音節認識結果の音節 3 連鎖を索引語として用い，音節認識結果を併用した実験を行った．音節認識単独での MAP 値は，0.287 であった．併用する単語認識の結果は，検索性能が高かった提案手法の検索質問拡張手法において，式 (20) の $\lambda = 0.76$ のときの結果を用いた．

図 4 に，式 (24) の単語認識と音節認識の補間係数 γ の値を $0 \leq \gamma \leq 0.02$ の間で 0.001 刻みに変化させたときの結果を示す．音節認識の結果を併用することで検索性能が向上していることが分かる．結果から，単語認識と音節認識を併用することは有効であることが分かった．

5.3.3 適合モデル

音節認識を併用する手法の実験を行った．音節認識結果の音節 3 連鎖を索引語として用いた．単独で音節認識を用いた際の最良の MAP 値は 0.274 であり，適合モデルを推定する Web 文書数は 40 文書，$\rho = 0.5$ であった．また，併用する単語認識の結果は，最も検索性能が高かった提案手法の検索質問拡張手法の結果を用いた．

図 5 に単語認識と音節認識の補間係数 ν の値を変化させたときの検索性能を示す．図から，$\nu = 0.36$ のとき最も性能が高く，わずかながら性能が向上した．ただし，併用の効果は小さいため，併用の方法などに検討の余地がある．

表 3 各手法の比較
Table 3 Comparison of proposed methods.

手法	Dry Run (MAP)	Formal Run (MAP)
ベクトル空間モデル	0.394	0.453
クエリ尤度モデル	0.464	0.500
適合モデル	0.476	0.501

5.4 各手法の比較

表 3 に，Dry run において各手法で最も性能が高かった場合のパラメータ設定による Dry run，Formal run における検索性能を示す．一般的に，ベクトル空間モデルよりも，クエリ尤度モデルや適合モデルといった確率的モデルに基づく方法が優れていることが分かる．この傾向は，一般的なテキスト検索と同様である．

Formal run における NTCIR-9 の公式の最高性能は MAP 値 0.427 であった [22]．今回の改良により，すべての手法においてその性能を上回った．

Dry run と Formal run の結果を比較すると，NTICR-9 のタスクは Formal run のほうがやさしいであろうことが想像される．しかし，手法間の精度の関係はほぼ同一であることから，Dry run で設定したパラメータに大きく左右されてはいない，すなわち非常にセンシティブでタスク変化に弱いパラメータ設定は存在しないであろうことが示唆される．しかし，ある程度のタスク依存性は存在し，実際に予備実験において，別のパラメータ設定で Formal run を実行した結果，表 3 よりも若干良い結果も得られた．今回の提案手法は音声認識に関しては，未知語率や誤認識率，言語的にはクエリの長さなどに影響を受けるものと考えられ，これらと頑健性（性能）がどのような関係にあるのかの分析は今後の課題である．

6. まとめ

本論文では，従来の 3 つの主な検索手法（ベクトル空間モデル，クエリ尤度モデル，適合モデルに基づく手法）に対し統一的な枠組みで改良を加えることで，音声ドキュメント検索における未知語や認識誤りに対処する手法を提案した．各検索手法に対し，未知語に対する対処として新たな検索質問拡張手法を提案した．また未知語と認識誤りに対し音節 3 連鎖を単語と同様に扱い検索し，スコアレベルで組み合わせることで，単語認識と併用する手法を提案した．

提案手法の有効性を NTCIR-9 の SpokenDoc タスクで評価し比較した．実験の結果，まず，各手法で Baseline となる手法と比較し，検索性能が向上することが明らかとなり，提案手法の有効性を示すことができた．なかでも特に，クエリ尤度モデルに基づく手法と，適合モデルに基づく手法では検索性能が高かった．提案手法は NTCIR-9 で公表されている公式の最高精度の結果を上回る結果を得た．

参考文献

[1] 秋葉友良：音声ドキュメント検索の現状と課題，SLP-82, Vol.2010, No.10, pp.1–8 (2010).

[2] Turunen, V.T.: Reducing the effect of OOV query words by using morph-based spoken document retrieval, *9th Annual Conference of the International Speech Communication Association*, pp.2158–2161 (2008).

[3] 岩見圭祐，藤井康寿，山本一公，中川聖一：音節 n-gram インデックスによる未知語の音声検索法の改善，情報処理学会研究報告，SDPWS，音声言語情報処理，Vol.2011, pp.1–6 (2011).

[4] Saraclar, M. and Sproat, R.: Lattice-based search for spoken utterance retrieval, *HLT-NAACL*, pp.129–135 (2004).

[5] Terao, M., Koshinaka, T., Ando, S., Isotani, R. and Okumura, A.: Open-Vocabulary Spoken-Document Retrieval Based on Query Expansion Using Related Web Documents, *9th Annual Conference of the International Speech Communication Association*, pp.2171–2174 (2008).

[6] 杉本樹世貴，西崎博光，関口芳廣：音声ドキュメント検索における Web ページを用いたドキュメント拡張の効果，情報処理学会研究報告，SLP，音声言語情報処理，Vol.2009, No.11, pp.1–7 (2009).

[7] 増村　亮，伊藤彰則ほか：確率的言語モデルに基づく音声ドキュメント検索のための Web を利用したモデル拡張の検討，情報処理学会研究報告，SLP，音声言語情報処理，Vol.2010, No.20, pp.1–6 (2010).

[8] Chen, B.: Word topic models for spoken document retrieval and transcription, *ACM Trans. Asian Language Information Processing* (*TALIP*), Vol.8, No.1, pp.1–27 (2009).

[9] Blei, D.M. and Lafferty, J.D.: Topic models, Text Mining: Theory and Applications, pp.71–93 (2009).

[10] Hofuman, T.: Probablistic latent semantic indexing, *Proc. 22nd Annual Internatiojnal ACM SIGIR Congerence on Research and Development in Information Retrieval*, pp.50–57 (1999).

[11] Salton, G. and McGill, M. (Eds.): *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).

[12] MacKay, D.J.C. and Bauman Peto, L.C.: A hierarchical Dirichlet language model, *Natural Language Engineering*, Vol.1, No.3, pp.289–308 (1995).

[13] Rocchio, J.J.: *Relevance feedback in information retrieval*, The SMART Retrieval System–Experiments in Automatic Document Processing, pp.313–323, Prentice Hall Inc. (1971).

[14] Ponte, J.: A Language Modeling Approach to Informatin Retrieval, PhD thesis, Dept. of Computer Science, University of Massachusetts, Amherst (1998).

[15] Lavrenko, V. and Croft, W.B.: Relevance based language models, *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.120–127 (2001).

[16] Lavrenko, V.: A Generative Theory of Relevance, Doctoral thesis, University of Mass., Amherst (2004).

[17] Ichikawa, K., Tsuge, S., Kitaoka, N., Takeda, K. and Kita, K.: Spoken document retrieval using both word-based and syllable-based document spaces with latent semantci indexing, *Proc. APSIPA ASC 2013*, 5 pages, (2013).

[18] Abdul-Jaleel, N. et al.: UMASS at TREC2004, *13th Text Retrieval Conference* (*TREC 2004*) *Notebook* (2004).

[19] Deerwseter, S., Dumais, S.T., Furnas, G.W. and Landauer, T.K.: Indexing by Latent Semantic Analysis, *Joural of American Society for Information Science*, pp.391–407 (1990).

[20] Iwata, K., Shinoda, K. and Furui, S.: Robust Spoken Term Detection Using Combination of Phone-Based and Word-Based Recognition, *Proc. Interspeech*, pp.2195–2198 (2008).

[21] Google: JSON/Atom Custom Search API (2013), available from ⟨https://developers.google.com/custom-search/json-api/v1/overview⟩.

[22] Akiba, T., Nishizaki, H., Aikawa, K., Kawahara, T. and Matsui, T.: Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop, *Proc. 9th NTCIR Workshop Meeting*, pp.223–235 (2011).

市川 賢

2012 年名古屋大学工学部電気電子・情報工学科卒業．2014 年同大学院情報科学研究科メディア科学専攻修士課程修了．

北岡 教英 （正会員）

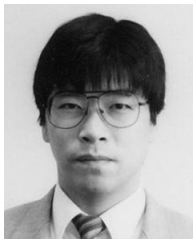
1994 年京都大学大学院工学研究科修士課程修了．1994 年（株）デンソー入社．2000 年豊橋技術大学大学院工学研究科博士後期課程修了．2001 年同大情報工学系助手．2003 年同講師．2006 年名古屋大学大学院情報科学研究科助教授．2007 年同准教授．2013 年徳島大学大学院ソシオテクノサイエンス研究部教授．2009 年 Nanyang Technological University visitingassociate professor．主として音声認識，音声対話，音声インタフェースに関する研究に従事．IEEE，ISCA，電子情報通信学会，日本音響学会，人工知能学会各会員．博士（工学）．

柘植 覚 （正会員）

1996 年徳島大学工学部知能情報工学科卒業．1998 年同大学大学院工学研究科博士前期課程知能情報工学専攻修了．2001 年同大学大学院工学研究科博士後期課程システム工学専攻修了．2000 年徳島大学工学部助手．2006 年徳島大学ソシオテクノサイエンス研究部講師．2010 年大同大学情報学部准教授．主として音声音楽情報処理，情報検索の研究に従事．日本音響学会，電気学会各会員．博士（工学）．

武田 一哉 （正会員）

1985 年名古屋大学大学院工学研究科修士課程修了．1985 年国際電信電話（株）入社．1986 年（株）ATR 国際電気通信基礎技術研究所．1990 年 KDD 研究所復職（この間 1988 年から 1989 年マサチューセッツ工科大滞在研究員），1995 年名古屋大学大学院工学研究科助教授，現在，同大学情報科学研究科教授．音声符号化，空間音響処理，音声情報処理等音声音響言語処理の研究に従事．日本音響学会理事，IEEE 会員．工学博士．



北 研二

1981 年早稲田大学理工学部数学科卒業．1983 年沖電気工業（株）入社．1987 年 ATR 自動翻訳電話研究所出向．1992 年徳島大学工学部講師．1993 年同助教授．2000 年同教授．2002 年同大学高度情報化基盤センター教授．2010 年大学院ソシオテクノサイエンス研究部教授．工学博士．自然言語処理，情報検索等の研究に従事．1994 年日本音響学会技術開発賞受賞．著書『確率的言語モデル』（東京大学出版会），『情報検索アルゴリズム』（共著，共立出版）等．