情報検索における“おもてなし”を実現するメディア処理技術

貞光 九月 †1 /島村 潤 †1 /入江 豪 †1
田良島 周平 †1 /吉田 大我 †1 /東中 竜一郎 †1
西川 仁 †1 /宮崎 昇 †1 /井島 勇祐 †1
中村 幸博 †2

NTTメディアインテリジェンス研究所 †1
NTTサービスエボリューション研究所 †2

本稿では，日々の活動におけるユーザの身の周りで個人をアシストするような，情報検索における“おもてなし”を実現するサービスイメージと，それを支える「画像による被写体識別技術」「人を理解し，自然に応答するための自然言語処理技術」「多様な話者・口調の合成音声を生成可能な音声合成技術」について紹介します．

情報検索における“おもてなし”

NTT研究所では日々のさまざまな活動シーンにおいて，ユーザ１人ひとりに向けて，きめ細やかで，利用者の属性に応じたサポートをするサービスの実現を目指しています．以下に，その具体的なイメージを紹介します．

■初めて見る商品の情報を利用者に合わせて提示するサービス

普段見慣れない料理や民芸品など，初めて見る身の周りの商品の情報をその利用者に合わせて提示してくれるサービスです（**図１**）．NTT研究所が開発した被写体識別技術を用いて，普段使っているスマートフォンやタブレットのカメラでかざした商品を高速に識別し，インターネット上に点在する関連情報と利用者の文化や嗜好などの個人属性を活用して，利用者に適したコンテンツを母国語で提供します．例えば，訪日外国人への日本料理や民芸品などの情報の母国語での提示，文化的背景や食品アレルギーなどにより食事に気を付けている方への料理の原材料の表示，買い物客への商品に関する評判情報の提示などのサービスが期待されます．

■ユーザに寄り添うぬいぐるみエージェントサービス

旅行やスポーツ観戦等のイベントなど，家族や仲間と過ごす楽しい時間，ふとバッグの中のスマートフォンを取り出して情報を調べたくなる場面はよくあります．しかし，それを手にした瞬間，家族や仲間の輪から切り離され，個人の世界に閉ざされてしまう，そんな経験はないでしょうか．

もしも，ユーザ間の何気ない会話の中からユーザの意図をとらえ，必要とされる情報をタイムリーに伝えることができれば，人の輪を崩すことなく，むしろ輪の中の新たな一員としてユーザへの情報提供を行うことができるでしょう．

NTT研究所の新たに開発したエージェントは，**図２**に示すように実体（例えばぬいぐるみや人形）を持ち，人の輪の中で物理的に共存します．そして，人の意図を理解したうえで，適

図１　初めて見る商品の情報を利用者に合わせて提示するイメージ

切な内容と量で発話文を生成，さらに多様な合成音声でしゃべることで，人の輪の中でも自然に振る舞うことが可能となります．

近い将来，さまざまな人の輪の中にエージェントがいる世界がやってくるかもしれません．

情報検索における“おもてなし”を支えるメディア処理技術

情報検索における“おもてなし”のサービスを実現するために，NTT研究所では「画像による被写体識別技術」「人を理解し，自然に応答するための自然言語処理技術」「多様な話者・口調の合成音声を生成可能な音声合成技術」の研究開発を推進しています．

■画像による被写体識別技術

ユーザに寄り添い，ユーザ状況に応じた情報を提供するためには，「コンピュータが，人と同じように世界やモノを識別・理解する」ことがキーポイントとなります．ここでは，人の目に相当するカメラの画像から，撮影された被写体を識別する被写体識別技術について紹介します．画像から被写体を識別するには，被写体に関する参照画像をあらかじめ準備しておく必要があります．撮影条件，環境条件，照明条件の違いに対応するため，通常は1つの被写体に対して多数の参照画像を準備する必要があり，その準備に大変な手間を要します．ここでは，この手間を大幅に削減する「3次元被写体識別技術」「クラウドデータ活用型被写体識別技術」について，それぞれ紹介します．

⑴　3次元被写体識別技術

「3次元被写体識別技術」は被写体が立体的な場合でも，少数の参照画像で高精度に識別することを可能にする技術です（**図3**）．本やCDなどの平面物と比べ，立体物は撮影方向によって画像上の見え方が大きく変わるため，従来はあらかじめ多数の参照画像を準備する必要がありました．本技術では，参照画像に対する相対的な撮影方向を自動推定することにより，ユーザが正面を意識せずに撮影した画像からでも，より頑健に立体物を識別でき

図2　ユーザに寄り添うぬいぐるみエージェントサービス

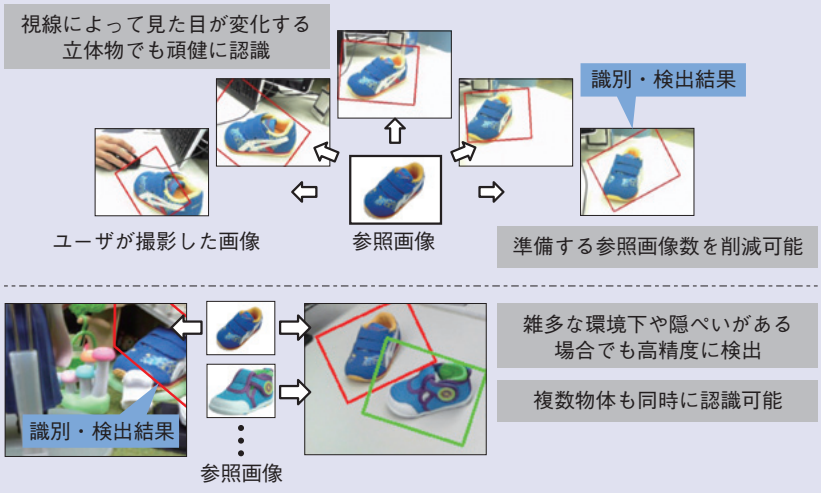


図3　3次元被写体識別技術の特長

ます．サービス事業者の立場では少数の方向からの画像を登録しておくだけで良いため，事前に用意する画像数を大幅に削減することができます．また本技術では，射影幾何学から導かれる立体物上での拘束条件を満たす特長的な点を画像からとらえることで，雑多な環境下や隠ぺいがある場合でも高精度に複数の物体を識別可能となっています．

⑵　クラウドデータ活用型被写体識別技術

参照画像は，被写体だけが正確にとらえられた，余計な背景のないものであることが望まれます．「クラウドデータ活用型被写体識別技術」（**図4**）では，独自の「被写体領域抽出技術」によって，この問題の解決を図っています．この技術では，識別したい被写体が写る複数枚の画像から，その被写体が存在する領域だけを特定して切り出すことができます（**図5**）．これを活用することで，手持ちの写真や，インターネットから取得した画像をまとめてクラウドに登録するだけで，手間を掛けずに被写体だけが写った参照画像を作成することができるようになります．被写体識別技術の導入・利用障壁を下げるだけでなく，膨大な参照画像を準備しやすくなるので，識別可能な被写体の種類を増加させることも可能になっていくと考えています．

大規模な参照画像を扱えるようになってくると，それをさばけるだけの高速な識別方式が不可欠です．NTT研究所では，「クロスモーダルハッシング」という独自のインデクシング技術の研究開発も進めています．この技術は，参照画像の内容を保存したごく短い符号（ハッシュ）に変換して索引化するもので，膨大な参照画像に対しても，索引を利用した高速な識別処理を実行できます．例えば，100万枚規模の参照画像に対して，これまでの被写体識別では7秒程度の処理時間を必要としていましたが，NTT研究所の技術ではこれを0.5秒以下で実行可能なレベルにまで到達しています．

今後さらに，画像から多数の被写体を高速・高精度に識別する技術へとブラッシュアップしていくとともに，ユーザ状況に応じて気の利いたアシストを行える使い勝手の良いサービスの実現を推進していきます．

■人を理解し，自然に応答するための自然言語処理技術

エージェントの自然な振る舞いのために，動作などの物理的機能も必要で

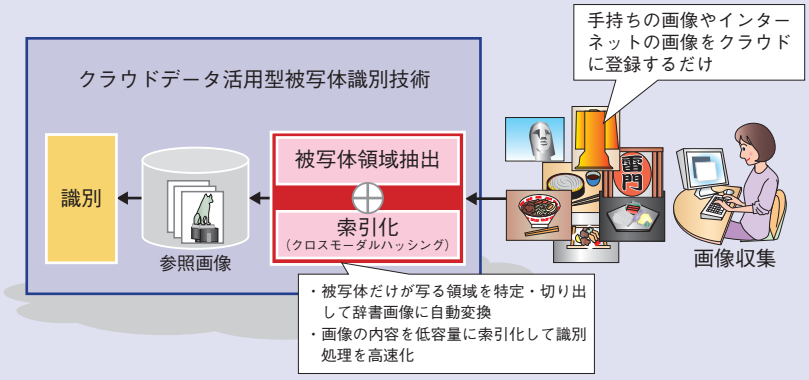


図4　クラウドデータ活用型被写体識別技術

図5　被写体領域抽出技術の特長

すが，エージェントが話す内容を生成するための自然言語処理技術も重要です．ここでは，①ユーザの発話意図理解技術，②自然な説明文生成のための自動要約生成技術，③対話全体を自然にまとめあげるための雑談対話技術，について紹介します．

⑴　発話意図理解技術

エージェントシステムの全体像を**図6**に示します．ここではエージェントが，祖父と孫という人の輪の中で，観光ガイド役を務める場面を例にとって説明を行います．

まずエージェントが行うべきは，ユーザの発話内容に対する意図理解です．コンピュータは人間の言語を理解することができないため，私たちは，人間の言語をコンピュータの言語（例えば，データベース問い合わせ言語）へと，一種の翻訳を行う技術を開発しています[4]．例えば，「いつできたんだろう」という文と，現在地などのメタ情報から，“s=電電寺，p=創建年，o=?”というコンピュータの言語に翻訳できれば，「電電寺がつくられたのは1252年です」と答えることが可能です．

⑵　自動要約生成技術

しかし，先述のように一言で回答を返すだけでは自然なエージェントとはいえません．私たちは適切に要約された説明文を回答に付加することで，自然さと賢さを向上する研究を行っています．私たちのアプローチは，Web上にすでに存在する，ある対象（電電寺）に関するテキスト情報を，説明としてふさわしいかたちに変換しエージェントに発話させ情報の提供を行うというものです．このとき，自動要約技術[5]が大きな役割を果たします．Web上のテキストを説明としてふさわしいかたちに変換する際には，自動要約技術を用いて元々のテキストの冗長性を除去した簡潔なテキストを生成し，またそれを話し言葉らしく変換することで，自然な説明を提供します．これにより，自然な説明を行うエージェントが低コストで実現できます．図6の右側は，Web上のテキストが話し言葉らしく変換される際の例です．例えば「寺伝によれば，禅僧野比武山が1252年に開眼供養を済ませたとあり，この年をもって創建年とする……」というテキストは，「1252年に野比武山によってつくられたといわれているよ」という簡潔で口語的な表現に変換することで，発話として利用することができます．

⑶　雑談対話技術

価値のある情報を対話の中で提供することにより，対話の質を向上させることができますが，一方で対話全体の自然性を向上させるには，話題の網羅性を高める必要があります．そこで必要となるのが雑談機能です．

雑談機能はアルゴリズムとして落とし込むことが難しく，これまで手作業によるルールで実装されてきました．しかし，ルールによる方法はコストが高く，話題の網羅性も低いという問題点がありました．そこで，NTT研究所ではインターネット上のテキストデータを言語処理技術によって対話知識とすることで，幅広い話題について自動的に雑談ができる対話システムを構築しました[6],[7]（**図7**）．システムは，現在の話題とシステムの発話意図に基づいて，述語項構造データ（主語や目的語などからなる文の基となる構

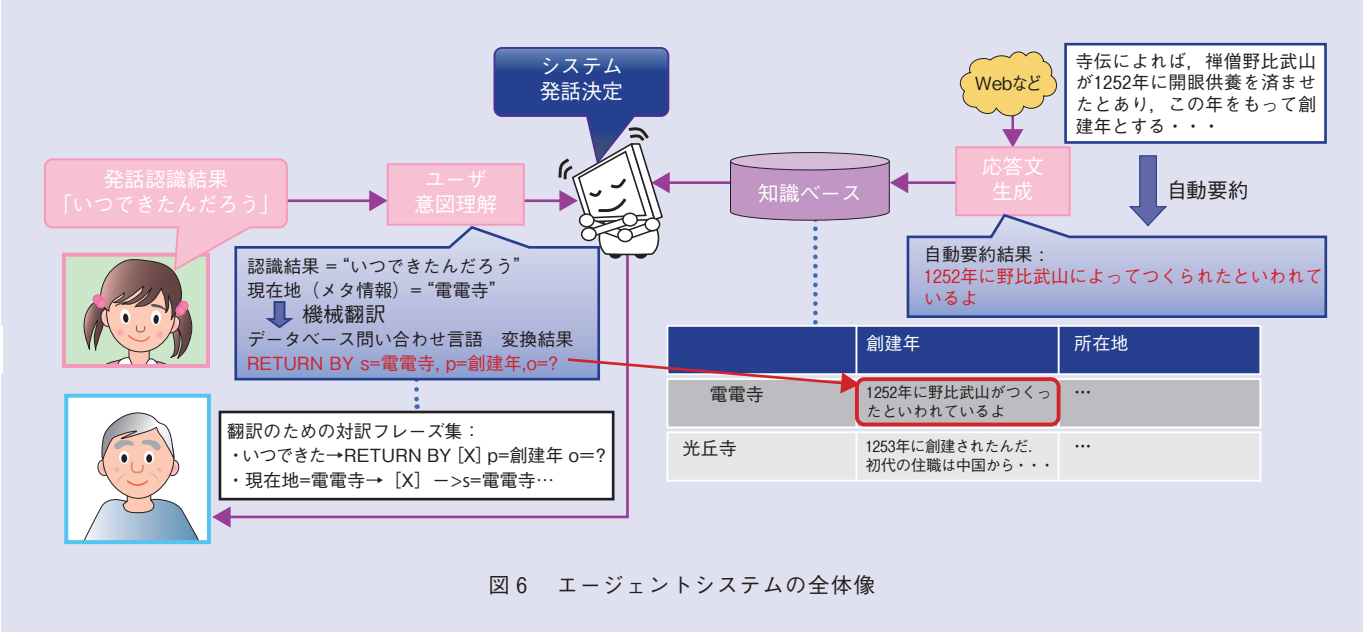


図6　エージェントシステムの全体像

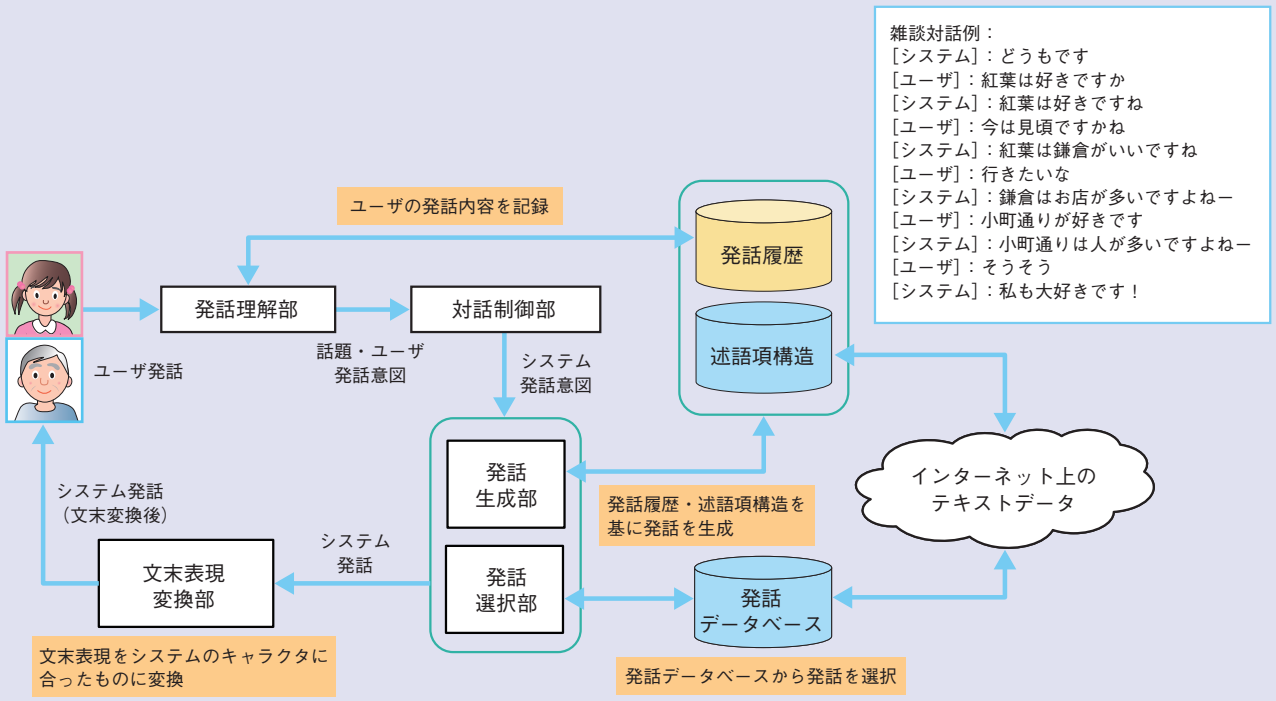


図 7　雑談対話技術の処理の流れ

造）から発話を生成したり，発話データベースから発話文を選択したりすることで，幅広い話題について応答を行います．文末表現変換機能により，システムのキャラクタに合った言葉づかいをすることも可能です．

このように，高度な対話エージェントを実現するためには，自然言語処理技術が必要不可欠です．今後も，深い言語理解と目的に応じた言語生成のための研究開発を推進していきます．

■多様な話者・口調の合成音声を生成可能な音声合成技術

これまで音声合成技術の主な利用用途は，音声での安否情報確認やコンタクトセンタの自動音声案内システムなどの情報提供サービスでしたが，近年の高性能なモバイル端末の普及に伴い音声対話エージェントなどへの利用用途が拡大しています．情報提供サービスは，音声で情報を正確に伝えることが目的であるため，音声合成に求められる機能として 1 名の話者が淡々と読み上げる口調の合成音声を生成できれば十分でした．一方，音声対話エージェントなどでは，エージェントのキャラクタに応じたさまざまな話者の合成音声，感情などに代表される場面に応じた口調の合成音声など，これまでとは異なる多様な話者・口調の合成音声が求められるようになってきています．ここでは，このような利用シーンを想定してNTT研究所が開発した，多様な話者・口調の合成音声の生成を可能とする，「ユーザデザイン音声合成技術」について紹介します．

ユーザデザイン音声合成の処理の流れを**図 8** に示します．本技術は任意の話者の音声から，その話者の音声の特徴を保持するモデルを学習する学習部と，学習したモデルなどを用いて合成音声を生成する音声合成部から構成されています．

学習部では，合成音声として再現したい話者の音声を収録し，収録した音声からその話者の音声の特徴（声質，口調）を表現するモデルを学習します．学習するモデルは，その話者の声質の情報を保持する話者モデル，声の高さや話すスピードといった話し方（口調）の情報を保持する口調モデルの 2 つのモデルから構成されています．

音声合成部では，学習した話者の話者モデル，口調モデルを用いて，合成したいテキストを音声に変換することで，その話者の声質，口調を持つ合成音声が生成されます．また，音声合成を行う際に，あらかじめ学習してある口調モデルの中から合成音声に付与したい口調，例えば執事風の口調や朗読風の口調など，に基づいて口調モデルを選択することで，その話者の声質を有しつつ指定された口調が付与された合成音声（例：Aさんの声質で朗読風口調の音声）を生成することも可能です．

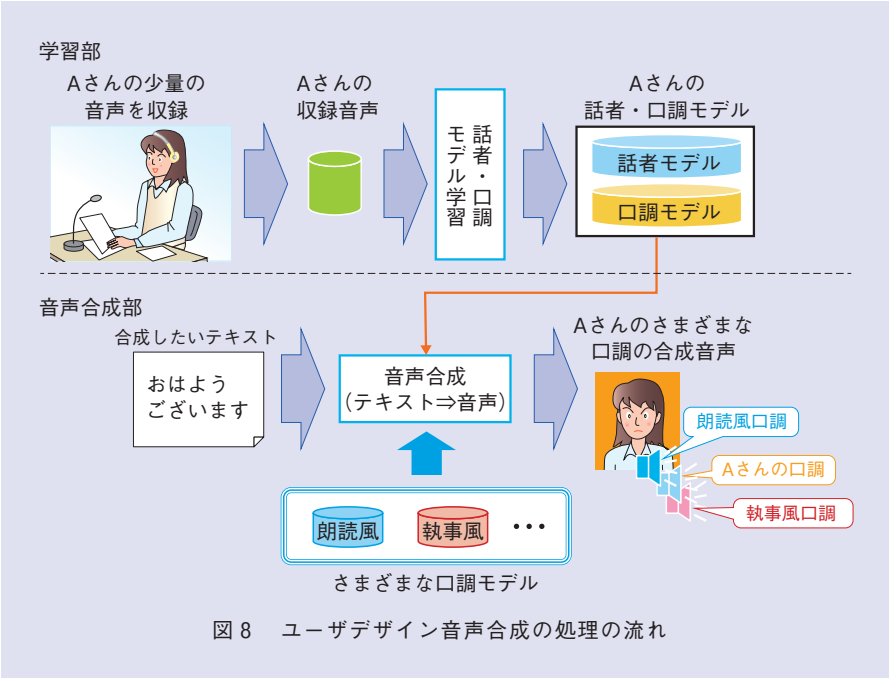


図 8　ユーザデザイン音声合成の処理の流れ

本技術の特長として，任意の話者の音声合成を実現する際に必要となるその話者の音声収録時間が短いということが挙げられます．これまでの音声合成技術では，十分な品質（話者性の再現性，合成音声の自然性）を持つ合成音声を生成するためには，数時間から数十時間におよぶ長時間の音声収録が必要なため，さまざまな話者の合成音声を作成することはコストなどの点で難しいという課題がありました．本技術では，必要となる音声収録時間を数十分から 2 時間程度と大幅に削減することで，エージェントのキャラクタに応じたさまざまな話者の合成音声の容易な作成を実現しています．

今後も，合成音声の自然性改善や話者性の再現性向上などの基本的な性能の向上に取り組むとともに，利用シーンに応じた適切な口調の実現など音声対話インタフェースに求められる機能の実現に向け研究開発を推進していきます．

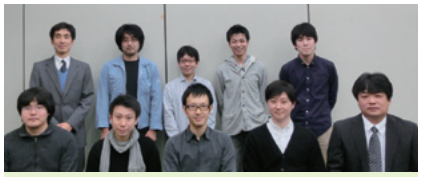
今後の展望

情報検索における“おもてなし”では，ユーザが日々の生活シーンで遭遇する課題に着目して，エージェントが働きかけていくことを目指しており，今回紹介した画像認識，言語処理，音声合成のさらなる技術開発がそれを支えていくと考えます．

今後は，潜在的な趣味嗜好や文化的な背景を理解したうえでの働きかけや，エージェントがユーザの輪の中に入り込んで，自然に情報を提供するという新しいスタイルを追求し，今までにないユーザとエージェントとの新しい関係が築けるようなサービスの実現を目指していきます．

■参考文献

(1) J. Shimamura, T. Yoshida, and Y. Taniguchi：“Geometric verification method to handle 3D viewpoint changes,” MIRU2014, OS3-4, Okayama, Japan, July 2014.
(2) 田良島・入江・新井・谷口：“領域マッチングに基づく高速なウェブ画像群物体コセグメンテーション，” 第42回画像電子学会年次大会，R4-2，2014.
(3) 入江・新井・谷口：“局所線形写像に基づくハッシング，” 信学論D-II，Vol. J97-D, No.12, pp.1785-1796，2014.
(4) 東中・貞光・内田・吉村：“しゃべってコンシェルにおける質問応答技術，” NTT技術ジャーナル，Vol.25，No.2，pp.56-59, 2013.
(5) H. Nishikawa, K. Arita, K. Takana, T. Hirao, T. Makino, and Y. Matsuo: “Learning to Generate Coherent Summary with Discriminative Hidden Semi-Markov Model,” COLING 2014, pp.1648-1659, Dublin, Ireland, August 2014.
(6) 東中：“雑談対話システムに向けた取り組み，” 言語・音声理解と対話処理研究会，Vol.70，pp.65-70，2014.
(7) R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano T. Makino, and Y. Matsuo：“Towards an open domain conversational system fully based on natural language processing,” COLING 2014, pp.928-939, Dublin, Ireland, August 2014.



（後列左から）宮崎　昇/ 東中　竜一郎/ 吉田　大我/ 島村　潤/ 田良島　周平
（前列左から）井島　勇祐/ 西川　仁/ 貞光　九月/ 入江　豪/ 中村　幸博

世の中にあふれる情報の中から，ユーザに必要な情報をより的確に返すことが“おもてなし”と称せる情報検索につながると考えます．NTT研究所で培ったメディア処理技術を，世界の人へ向けた“おもてなし”サービスとして具現化していきます．

◆問い合わせ先

NTTメディアインテリジェンス研究所
第一推進プロジェクト
TEL　046-859-5161
FAX　046-855-3495
E-mail　ozawa.shiro@lab.ntt.co.jp