DBSJ Journal, Vol.8, No.3
December 2009

課題解決型ハイパーリンク自動生成方式の開発とコンタクトセンターへの適用

Development of Problem-solving-oriented Automatic Hyperlink Creation Framework and its Application on Contact Center

立石 健二♥　細見 格♦, 久寿居 大♦

Kenji TATEISHI, Itaru HOSOMI, Dai KUSUI

本稿では，コンタクトセンター業務におけるオペレータの情報収集支援を目的とした課題解決型のハイパーリンク自動生成方式を提案する．ハイパーリンク自動生成システムは，過去の問合せ事例やメンテナンスマニュアル等，問合せ回答に用いる様々な情報源間にリンクを生成し，問合せ回答のための情報収集時間と情報収集漏れを削減する．従来のリンク生成方式は，文書からキーワードに該当する単語列を抽出し，キーワードが共通する異なる情報源の文書間にリンクを生成するが，参照先文書の絞込みとキーワードの抽出性能向上の課題があった．これらを解決するため，本稿では課題解決型のハイパーリンク生成方式と，同方式における新しいキーワード抽出方式を提案する．

In this paper, we propose a problem-solving-oriented hyperlink creation framework for information gathering on contact center. Hyperlink creation systems connect many information resources used for operator's question-answering such as case documents and maintenance manuals, and reduce both the time of information gathering and the lack of information to be needed. The current link creation approach extracts topic keywords from documents, and creates hyperlinks among documents of different information resources sharing the same keyword. However, it has two issues: to limit link-to documents effectively and to improve topic keyword extraction performance. In order to solve them, we propose a problem-solving-oriented hyperlink framework and a new keyword extraction method on its framework.

1. はじめに

1.1 背景

コンタクトセンターの窓口業務効率化は企業活動における重要な課題となっている．主要な課題として回答の正確性向上と回答時間の短縮があり，これらは顧客満足度向上やコスト削減の効率化に密接に関係している．例えば，製品保守の窓口では保守要員からの製品の故障の問合せに対し，オペレータはその原因を予想し，被疑部品を保守要員に伝える．迅速な回答は迅速な修理につながり顧客満足度に影響し，正確な被疑部品の特定は保守コストの低減に貢献する(部品交換を必要最小限にとどめる)．

正確かつ迅速な回答のためには，社内外に存在する様々な情報源を用いて十分な情報収集が必要となる．例えば，製品保守の窓口のオペレータは，保守要員からの問い合わせの際に，過去の問合せ履歴データや，ユーザ/メンテナンスマニュアル，製品故障通知，公式Webサイト等，様々な情報源を参照して回答する．しかし，経験の浅いオペレータにとってこの問合せ回答のための情報収集は必ずしも容易ではない．その理由として，回答に必要な情報源が分散していること，及び，問合せの種類に応じて参照すべき情報源が異なることがあげられる．例えば，オペレータは問合せを受けると関連する過去の問い合わせ履歴を調査するが，その詳細はメンテナンスマニュアルで確認する必要がある．また，メンテナンスマニュアルに記載されていない新規の故障情報は別の情報源(製品故障通知等)を参照する必要がある．さらに，同一の企業の製品保守であってもパーソナルコンピュータとプリンタとでは全く異なる情報源を参照する必要がある．したがって，オペレータは問合せの種類毎にどの情報源のどの箇所にどのような内容が記載されているかをあらかじめ把握しておく必要があり，このことが情報収集の難しさの原因になっている．

この状況を改善するためには，情報源の再構築が最も直接的な対応方法であるが，膨大な労力が必要となる．そのため，情報源は分散したまま，オペレータが効率的に情報の収集をする仕組みが現実的である．この支援の方向性として，大きく「分散した複数の情報源を横断的に検索する横断検索」と「情報源の特定の項目間を関連付けるハイパーリンク生成」の2つがある．前者は，物理的に情報源は分散しているが，一回の問合せで参照すべき情報は一箇所に記載されている場合に適切である．一方後者は，物理的に情報源が分散しており，一回の問合せ(オペレータが電話を受けてから電話を切るまで)に参照すべき情報も複数個所に分散している場合に適切である．企業の保守窓口の現状には後者がより近いと考え，本稿ではハイパーリンク生成に焦点を当てる．

本研究が想定するハイパーリンク自動生成システムの動作イメージを図１に示す．本システムは，例えば問合せ履歴データとメンテナンスマニュアルといった異なる情報源間にハイパーリンクを自動的に生成する．ハイパーリンクによる情報収集への具体的な支援効果は以下の２つである．一つ目は，関連事項の記載箇所を探す時間を削減できることである．例えば，過去の問合せ事例を参照してその対応方法に記載された「カートリッジ交換」を問合せ元の保守要員に伝えたとする．この状況で「カートリッジ交換」の具体的な手順が知りたいと追加質問を受けたとき，オペレータはリンクをたどるだけで手順を記載した別の情報源の文書へ到達できる．二つ目は，関連事項の存在に関する気づきをオペレータに与え，情報の収集漏れを削減できることである．例えば，同様な状況で「カートリッジ交換」にアンカーテキストが存在し，その参照先文書として「保守通知書：カートリッジ交換時の注意事項」が関連付けられていれば「カートリッジ交換」の際には読まなければならない情報が別の文書に存在することをオペレータが認識し，その記載内容を保守要員に通知できる．

なお，本システムでは全ての情報にWebを通してアクセス

♥ 正会員 NEC 共通基盤ソフトウェア研究所
k-tateishi@bq.jp.nec.com
♦ 非会員 NEC 共通基盤ソフトウェア研究所 {kusui@ct, i-hosomi@ay}.jp.nec.com

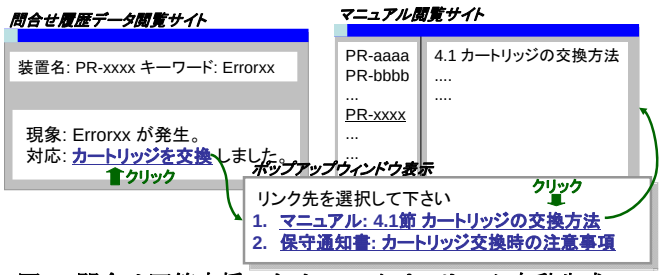

図 1 問合せ回答支援のためのハイパーリンク自動生成システム
Fig.1 Automatic Hyperlink Creation System for Supporting Question-answering on Contact Center

できることを想定する．すなわち，一つの情報源とは一つのWebサイトを指し，メンテナンスマニュアルのWebサイトや過去の問合せ事例のWebサイトが個別に存在する環境を想定する．

1.2 従来技術

ハイパーリンク自動生成システムは，リンク対象文字列(アンカーテキストを設定する文字列)により3種類に分類できる．一つ目はリンク対象文字列をキーワードとする形態である．従来研究では，辞書の説明文の中の他のエントリー部分をリンク対象文字列とする方法[1]やニュース内の記事の人名や土地名をリンク対象文字列としてWeb検索結果を表示する方法[2]，電子図書の関連する語句間にハイパーリンクを生成する研究[3]がある．次に，リンク対象文字列の単位をパラグラフとする形態がある．製品マニュアルの段落から関連するリファレンスマニュアルの段落へのリンクを生成する方法が提案されている[4]．この形態では，パラグラフそのものをリンク文字列とするのではなく，パラグラフの下側に参照先の文書内容の概要を示すリンク文字列を表示する．最後に，リンク対象文字列を利用者自身がマウス等で指定する形態がある．TV画面でWebを閲覧する場合に，利用者がリモコンポインターで指定した位置をキーワードとして検索する方法が提案されている[5]．

本研究ではこの中でリンク対象文字列をキーワードとする形態を採用する．これは，問合せ回答のための情報収集では，操作(例.カートリッジ交換)，障害現象(例．紙詰まり)，部品，エラーコード等のキーワードを起点として異なる情報源を参照する要求が多いと考えられるためである．また，リンク対象文字列をアンカーテキストで区別することによって，オペレータに対し関連する重要文書が存在するという気づきを与え，情報収集漏れを削減するという効果を実現できるからである．リンク対象文字列をキーワードとするハイパーリンク生成システムでは，文書からキーワードに該当する単語列を抽出し，キーワードが共通する文書間にリンクを生成する．図2の文書1を中心に説明すると，文書1に含まれるキーワード「コール50」の参照先文書は，文書1以外でキーワード「コール50」が含まれる文書2，文書3である．

キーワード抽出方式は，辞書やルールを用いて事前定義されたカテゴリのキーワードを抽出する方式[6]と，統計的に文書のトピックとなるキーワードを抽出する方式に分かれる．辞書や抽出ルールに基づく方式は，十分な事前知識を与えられれば高精度なキーワード抽出を実現できるという利点を

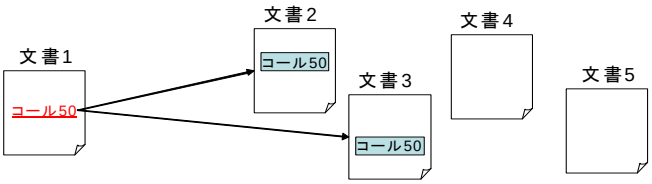

図 2 キーワード抽出に基づくリンク生成方式
Fig.2 Hyperlink Creation Method Based on Keyword Extraction

持つ一方で，統計的手法は，辞書やルールの作成やメンテナンスが不要で，運用コストが低いという利点を持つ．また，問合せ回答のための情報収集で必要となるキーワード(操作方法，障害現象，部品名，エラーコード名等)は，文書に記載された問題あるいは対応方法の中心的役割を果たす語と考えられるため，統計的手法よっても抽出可能である．したがって実用上は，容易に思いつくキーワードは辞書やルールで用意し，不足を補う(網羅性を高める)ために統計的手法で補完するという組み合わせが良い．

統計的手法に基づくキーワードの抽出方式としてtf/idf[7]が良く知られている．tf/idfは文書のトピック(文書の内容を代表する単語列)をキーワードとする方法である．tfとidfはそれぞれ文書内で出現回数が多い単語列，出現文書数が少ない単語列を重視する指標であり，tf/idfはこの2つの指標の積により単語列の重要度を計算し，各文書で重要度の高い単語列をキーワードとする．その他の統計的手法として，中川らは着目する単語列を含む複合名詞の頻度あるいは異なり数が多い語をキーワードとする方式を提案している[8]．類似した方法としてC-Valueも知られている[9]．久光らは，着目する単語列を含む文書の単語列の出現分布と，全文書の単語列の出現分布の異なりの度合いを用いた方式を提案している[10]．

1.3 従来技術の課題

このような従来システムの課題として，(1)参照先文書の絞込みと(2)キーワードの抽出性能向上がある．(1)参照先文書の絞込みについては，従来研究[1,3]では，同一のキーワードを含む文書が多く存在する場合に，利用者に提示する参照先文書を絞り込む方法に関しては言及されていない．無論，図1のインターフェースのように複数を参照先候補として提示する方法も可能であるが，この場合でも利便性の観点から提示できる候補は少数に選別されるべきである．参照先候補が多いとその中から利用者が取捨選択する負担が大きくなるためである．

(2)キーワードの抽出性能向上に関しては，tf/idf は単語列の重要度を測るための必要条件ではあるが十分条件ではない．tf と idf はそれぞれ文書内で出現回数が多い単語列，出現文書数が少ない単語列を重視する指標であるが，実際には，単語列の文書内の出現回数の順位が中位程度の(極端に少なくはないが多くもない)場合，あるいは，単語列の出現文書数の順位が中位程度の(極端に多くはないが少なくもない)場合でもキーワードとすべき重要な単語列は存在する．高精度のキーワード抽出のためには，より多くの指標を組み合わせて重要度を測ることが重要であると考える．

1.4 研究の目的

本研究では，上記の課題を解決するため，オペレータの問合せ回答に結びつく情報へのリンクを生成する課題解決型のハイパーリンク生成方式を提案する．提案方式では，全ての情報源間にリンクを生成するのではなく，オペレータの典

型的な情報参照プロセス上に隣接する情報源間にリンクを生成する．オペレータの典型的な情報参照プロセス上に隣接する情報源間とは，「過去の問合せ事例を閲覧した後には，メンテナンスマニュアルを閲覧する」といった，多くのオペレータが現実に採用している参照順序の前後段に位置する 2 つの情報源を表す．これにより，参照先文書の適切な絞込みが可能となる．

また，課題解決型ハイパーリンク生成における新しいキーワード抽出方式を提案する．提案方式は，典型的な情報参照プロセス上に隣接する情報源間に出現するリンクの特徴を用いてキーワードとなる単語列を抽出する．従来方式(tf/idf)と，リンクの特徴を用いた提案方式を組み合わせることにより，文書のトピックとなる重要な単語列をより高精度に抽出する．従来方式と提案方式は独立した手法であり，tf/idf 以外のキーワード抽出方式に提案方式を適用することも可能である．

上記の 2 つの提案方式の詳細は，2 章，および，3 章で説明する．また，4 章でキーワード抽出方式の有効性を評価する．製品保守の窓口業務で使用する文書を用いて提案方式を評価したところ，従来方式(tf/idf)と組み合わせることにより有効なリンクを多く検出できることが明かになった．5 章でまとめる．

2. 課題解決型ハイパーリンク生成

提案する課題解決型のハイパーリンク生成方式は，オペレータの典型的な情報参照プロセスを明らかにし，そのプロセス上に隣接する情報源間にリンクを生成する(図 3 参照)．コンタクトセンターのオペレータは，問合せ回答のための情報の収集順序を経験的に保有している場合が多く，その順序に従ったリンクを生成することで，参照先文書を適切に絞込めると期待できる．

典型的な情報参照プロセスは，オペレータが現実に参照元/参照先として頻繁に利用する情報源を分析することで明らかにする．これは，業務の観察とオペレータへのヒアリングにより可能となる．オペレータの業務を観察することにより，例えば「オペレータは問合せを受けると，まず過去の問合せ履歴を閲覧し対応方法の概要を把握し，その詳細をメンテナンスマニュアルで確認する」という仮説を立てることが可能である．この仮説が正しいかどうかは，実際のオペレータにヒアリングすることで確認できる．このとき，2 つの情報源(問合せ履歴とメンテナンスマニュアル)は典型的な情報参照プロセスの順序をなしているといえる．

このように本稿では，コンタクトセンターの問合せ回答というタスクにおける典型的な情報参照プロセスを観察とヒアリングという手作業で作成する．一方，コンタクトセンター以外の異なるタスクに課題解決型のハイパーリンク生成方式を導入する場合は，情報源の範囲が特定できずこの方法による作成が困難な場合もあり得る．このようなタスクにおいては，利用者の Web ブラウザの閲覧履歴によりプロセスを作成する方法が考えられる．この場合，閲覧履歴から多くの利用者が情報源をどの順番で閲覧しているかを分析する．閲覧履歴によりシステムが典型的な情報参照プロセスを作成する方法に関しては今後の課題とする．

以下，オペレータの典型的な情報参照プロセス上に隣接する 2 つの情報源を単純に「適切な情報源ペア」と呼ぶ．図 4 に例を示す．適切な情報源ペアにおけるリンクの生成元の情報源を「参照元情報源」，リンクの生成先の情報源を「参照先情報源」と呼ぶ．つまり，本研究では参照元情報源から参照先情報源へのリンクを生成する．また，参照元情報源に属する文書を参照元文書，参照先情報源に属する文書を参照先文書と呼ぶ．

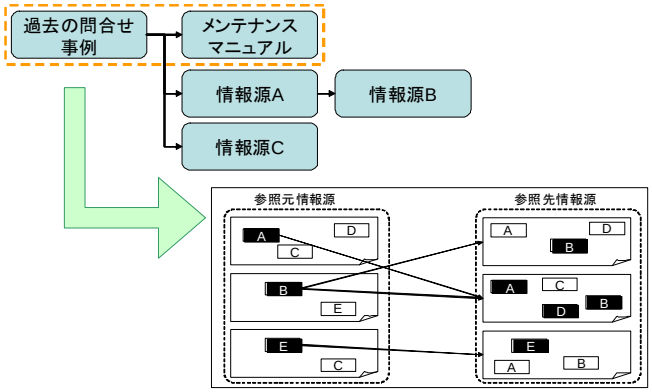

図 3 課題解決型ハイパーリンク生成方式
Fig. 3 Problem-solving-oriented Hyperlink Creation Framework

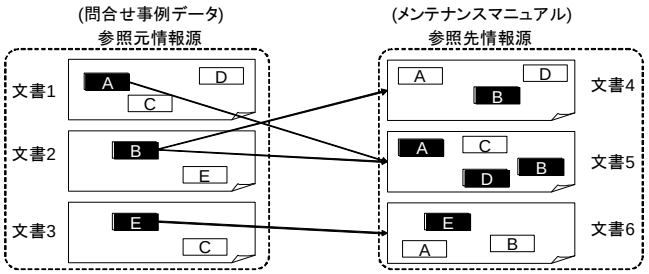

図 4 適切な情報源ペアに生成するハイパーリンクの例
Fig. 4 An Example of Hyperlinks created on a Pair of Proper Information Resources

リンク生成の基本手順は以下の通りである．まず双方の情報源の文書を形態素解析により単語列に分割する．次に，文書に含まれる単語列の中からキーワードとなる単語列を選択する．最後に，参照元文書のキーワードから同一のキーワードを含む参照先文書へのリンクを生成する．図 4 では，A,B,C 等のアルファベットが文書に含まれる単語列を表し，色つきが単語列の中から選択したキーワードである．例えば，参照元情報源の文書 1 のキーワード A は参照先情報源の文書 5 でもキーワードとなっている．そこで，文書 1 のキーワード A から文書 5 へのリンクを生成する．

3. 課題解決型ハイパーリンク生成のためのキーワード抽出方式

3.1 方式概要

提案するキーワード抽出方式は，適切な情報源ペアにリンクを生成する方式である．提案方式は，「正確なキーワード抽出に基づいて適切な情報源ペアに対して生成したリンクには一定の特徴がある」という仮説に基づく．このリンクの特徴については 3.2 節で説明するが，端的には図 5 の(z)の形態のリンクを表す．この仮説を正しいとした場合，「特徴と一致するリンクを生成する単語列が，正しいキーワードの可

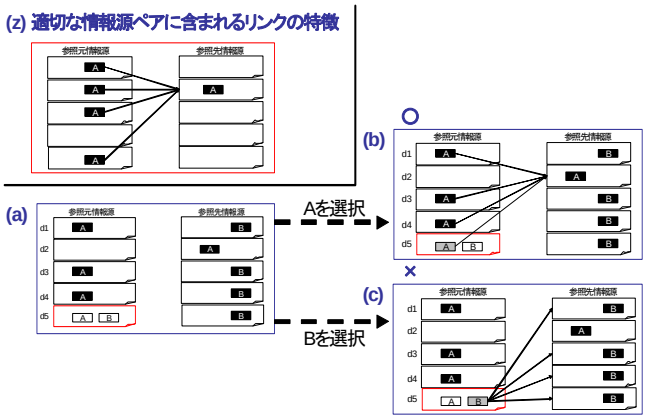

図 5 課題解決型ハイパーリンク生成のためのキーワード抽出方式概要

Fig. 5 An Overview of Keyword Extraction Method on Problem-solving-oriented Hyper Creation Framework

能性が高い」とする提案方式が導ける．適切な情報源ペアは 2 章で説明したように観察とヒアリングで選択する．したがって，提案方式は，コンタクトセンターの問合せ回答のような適切な情報源ペアを観察とヒアリングで選択できるタスクにおいて有効な方式である．

提案方式の動作例を図 5 を用いて説明する．図の(a)は適切な情報源ペアの例である．この，単語列 A と単語列 B を含む文書 d5 からキーワードを抽出する場合で説明する．なお，他の文書のキーワードは既に抽出されているとする．このとき，単語列 A をキーワードとして選択すると図の(b)のリンクが生成される．一方，単語列 B をキーワードとして選択すると図の(c)のリンクが生成される．今，図の(z)が前提とするリンクの特徴であるので，「特徴と一致するリンクを生成する単語列 A が単語列 B よりも正しいキーワードの可能性が高い」ことになる．

3.2 リンクの特徴

ここで，適切な情報源ペアに含まれるリンクの特徴を図 6 を用いて説明する．2 種類の特徴がある．

一つ目は，適切な情報源ペアでは，一つの参照先文書は多くの参照元文書のキーワードから参照されるという特徴である(リンクの特徴 1)．図 6 の例では，1 つの参照先文書は 4 つの参照元文書のキーワード A から参照されている．適切な情報源ペアはオペレータが現実に参照元/参照先として頻繁に利用している情報源ペアであるため，そこには潜在的に多数のリンクが含まれるはずである．このとき，平均的には 1 つの参照先文書は多くの参照元文書のキーワードから参照されることになる．

二つ目は，適切な情報源ペアでは，参照元文書のキーワードに対する参照先文書は少数に限定されるという特徴である(リンクの特徴 2)．図 6 の例ではキーワード A の参照先文書は 1 つである．現実に参照先として利用する機会が多い適切な情報源ペアの参照先情報源は，その利便性も一定の水準を満たしているはずである．また，利便性が高い文書セットでは，特定の話題に関する内容が分散せずに集約して(体系的に)記載されているはずである．

3.3 処理手順

提案方式は，下記の 3 つの手順で実行する．

1. 形態素解析を用いて文書を単語列に分割

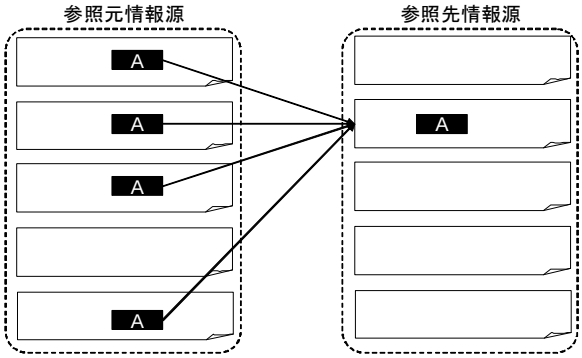

図 6 適切な情報源ペアに含まれるリンクの特徴

Fig. 6 Features of Hyperlinks contained on a Pair of Proper Information Resources

2. 従来方式(tf/idf)を用いて文書毎に単語列の重要度を計算し，重要度が高い上位 M%を仮のキーワードとする
3. 仮のキーワードにより生成したリンクとリンクの特徴との間の整合性に基いて単語列の重要度を補正し，重要度が高い上位 M%をキーワードとする

処理 3 で用いる単語列の重要度計算式を以下に示す．情報源 D の文書 d に含まれる単語列 t の重要度 S(D, d, t)は，2 つの項から構成される．第 1 項の BS は Base Score の略であり，従来方式(tf/idf)による重要度を表す．式の第 2 項が重要度の補正項である．

$$(式1)\quad S(D,d,t) = \mathrm{BS}(D,d,t) * \log_2(\mathrm{sdf}(D_{FROM}) + 1)$$

$$(式2)\quad S(D,d,t) = \mathrm{BS}(D,d,t) * \log_2\left(\frac{\mathrm{sdfmax}(D_{TO})}{\mathrm{sdf}(D_{TO},t)} + 1\right)$$

ここで(式 1)の D_{FROM} は参照元情報源を，$\mathrm{sdf}(D_{FROM}, t)$は D_{FROM} において t が仮のキーワードとなった文書数を表す．(式 1)は，参照元情報源で t が仮のキーワードとなった文書数が多いほど大きな値となるため，単語列 t とリンクの特徴 1 との整合性を考慮した補正項であるといえる．

同様に(式 2)の D_{TO} は参照先情報源を，$\mathrm{sdf}(D_{TO}, t)$は D_{TO} において t が仮のキーワードとなった文書数を表す．$\mathrm{sdfmax}(D_{TO})$は $\mathrm{sdf}(D_{TO}, t)$の D_{TO} における最大値を表す．(式 2)は，参照先情報源で t が仮のキーワードとなった文書数が少ないほど大きな値となるため，単語列 t とリンクの特徴 2 との整合性を考慮した補正項であるといえる．

上記の式において，参照元情報源の文書からキーワード抽出する際に(式 1)を適用すると，参照先情報源のどの文書においてもキーワードとならない単語列に大きな重要度が与えられる可能性がある．(式 1)の補正項は適切な情報源ペアに含まれるリンクの特徴から導かれた指標であるので，このようなリンクにならないキーワードを抽出すると有効性が失われる．参照先情報源の文書からキーワードを抽出する際に(式 2)を適用する場合も同様である．したがって実用上は，図 7 のような条件分岐を用いた計算式を用いる．

4 評価実験

本章では，課題解決型のハイパーリンク生成におけるキーワード抽出方式の有効性を評価する．なお，2 章で述べたように本稿では，適切な情報源ペアを観察とヒアリングにより

1. 参照元情報源から(リンクの特徴1)を用いて抽出する場合

$$\text{IF}\,(\text{sdf}\,(D_{FROM}, t) > 0 \wedge \text{sdf}\,(D_{TO}, t) > 0)\ \text{THEN}$$
$$\text{S}(D_{FROM,}\, d, t) = \text{BS}(D_{FROM,} d, t) * \log_2(\text{sdf}\,(D_{FROM}, t) + 1)$$
ELSE
$$\text{S}(D_{FROM,}\, d, t) = \text{BS}(D_{FROM,} d, t)$$

2. 参照元情報源から(リンクの特徴2)を用いて抽出する場合

$$\text{IF}\,(\text{sdf}\,(D_{TO}, t) > 0)\ \text{THEN}$$
$$\text{S}(D_{FROM,}\, d, t) = \text{BS}(D_{FROM,} d, t) * \log_2\left(\frac{\text{sdfmax}(D_{TO})}{\text{sdf}(D_{TO}, t)} + 1\right)$$
ELSE
$$\text{S}(D_{FROM,}\, d, t) = \text{BS}(D_{FROM,} d, t)$$

3. 参照先情報源から(リンクの特徴1)を用いて抽出する場合

$$\text{IF}\,(\text{sdf}\,(D_{FROM}, t) > 0)\ \text{THEN}$$
$$\text{S}(D_{TO,}\, d, t) = \text{BS}(D_{TO,} d, t) * \log_2(\text{sdf}\,(D_{FROM}, t) + 1)$$
ELSE
$$\text{S}(D_{TO,}\, d, t) = \text{BS}(D_{TO,} d, t)$$

4. 参照先情報源から(リンクの特徴2)を用いて抽出する場合

$$\text{IF}\,(\text{sdf}\,(D_{FROM}, t) > 0 \wedge \text{sdf}\,(D_{TO}, t) > 0)\ \text{THEN}$$
$$\text{S}(D_{TO,}\, d, t) = \text{BS}(D_{TO,} d, t) * \log_2\left(\frac{\text{sdfmax}(D_{TO})}{\text{sdf}(D_{TO}, t)} + 1\right)$$
ELSE
$$\text{S}(D_{TO,}\, d, t) = \text{BS}(D_{TO,} d, t)$$

図 7 重要度計算式の詳細
Fig. 7 Calculation of Term Importance Value

選択する.

4.1 対象文書

コンタクトセンターの問合せ回答で工業製品の保守部門のオペレータが実際に利用している下記の2種類の適切な情報源ペアを対象とした.

[実験データ 1] 問合せ履歴データ→メンテナンスマニュアル

参照元情報源である問合せ履歴データは，過去にオペレータが受け付けた問合せ内容のログを，質問内容である障害の現象と指示内容である対応方法に分けて記録したものである．参照先情報源であるメンテナンスマニュアルは，製品毎に機能や部品，操作手順を説明した文書である．問合せ履歴データの 1 つの事例を 1 文書とし，メンテナンスマニュアルの 1 節を 1 文書とした．この 2 つの情報源が適切な情報源ペアであることは，業務の観察とオペレータへのヒアリングにより確認した．オペレータは問合せを受けると，まず問合せ履歴データを閲覧し対応方法の概要を把握し，その詳細をメンテナンスマニュアルで確認するという参照順序を持つ.

なお，問合せ履歴データとメンテナンスマニュアルは製品毎に分かれている．異なる製品の情報源同士は適切な情報源ペアにならないため，適切な情報源ペアを製品毎に設定した．「プリンタ A」「プリンタ B」の 2 種類のプリンタを評価に用いた．「プリンタ A」に関する問合せ履歴データは 289 文書，メンテナンスマニュアルは 243 文書であった．また，「プリンタ B」の問合せ履歴データは 318 文書，メンテナンスマニュアルは 269 文書であった.

[実験データ 2] 問合せ履歴データ→保守事前通知書

参照元情報源である問合せ履歴データは上記と同一である．参照先情報源である保守事前通知書は，製品保守の際に保守要員が製品に関して注意すべき内容を記載した文書であり，1 つの内容を 1 文書とした．この 2 つの情報源が適切な情報源ペアであることは，業務の観察とオペレータへのヒアリングにより確認した．オペレータは問合せを受けると，過去の問合せ履歴データを閲覧し，対応方法の概要を確認し，その対応方法(例．カートリッジ交換)に関して注意すべき重要な内容が存在するかを保守事前通知書で確認するという参照手順を持つ.

問合せ履歴データと保守事前通知書は製品毎に分かれている．本実験では，「サーバ A」「サーバ B」の 2 種類のサーバマシンを評価に用いた．「サーバ A」に関する問合せ履歴データは 375 文書，保守事前通知書は 63 文書であった．「サーバ B」に関する問合せ履歴データは 420 文書，保守事前通知書 46 文書であった.

4.2 重要度計算手順

比較対象とする従来方式によるキーワード抽出方式は，tf/idf をベースとした下記の重要度計算方式 BS を用いた．tf(D, d, t)は，情報源 D の文書 d における単語列 t の出現回数を，df(D,t)は D における t の出現文書数を，|D|は D の総文書数を，length(t)は t の長さ(バイト数)をあらわす.

$$\text{BS}(D, d, t) = \text{tf}\,(D, d, t) * \log_2 \frac{|D|}{\text{df}\,(D, t)} * \log_2(\text{length}(t))$$

一方，提案方式は 3.3 節の流れでキーワードを抽出する．まず，処理 1 では形態素解析ツールとして ChaSen[11]を用いて文書を単語に分割し，「名詞」「未知語」の任意の組合わせの連続を単語列とした．この時，複数の単語から構成される単語列は，その部分単語列も単語列とした．処理 2 の仮のキーワード抽出に用いた重要度計算方式は比較対象と同様の BS であり，各文書から重要度の高い上位 5%の単語列をキーワードとして選択した. 処理(3)のキーワード抽出に用いた重要度計算方式は，[実験データ 1]の対象文書に関しては図 7 の 1 番目と 2 番目の計算式を用いて参照元情報源からキーワードを抽出した．1 番目がリンクの特徴 1，2 番目がリンクの特徴 2 を用いた重要度計算方式である．また，[実験データ 2]の対象文書に関しては 3 番目と 4 番目の計算式を用いて参照先情報源からキーワードを抽出した．3 番目がリンクの特徴 1,4 番目がリンクの特徴 2 を用いた重要度計算方式となる.

キーワードは相対的に長い単語列が多いことが経験的に知られており[9]，この経験則を 2 種類の方法で導入した．一つ目の方法は，上記の BS の length 項の追加である．二つ目の方法は，部分一致の可能性の高い単語列の変換である．最終的に選択したキーワードがその文書内において他の単語列の部分単語列としてのみ出現する場合を考える．例えば，選択したキーワードが「イメージ」で，この単語列が「イメージ」が含まれる文書内で「イメージ」単独では出現せず「イ

メージドラム」の部分単語列としてのみ出現する場合が該当する．上記経験則に従えば，「イメージドラム」が正解で「イメージ」はその部分単語列である可能性が高い．そこで，選択したキーワード A がその文書内で別の単語列A'の部分単語列としてのみ出現する場合は，部分一致の可能性が高いとして選択した A を A'に変換した．

4.3 評価方法

［実験データ 1］では参照元情報源からランダムに 200 文書を選択し，［実験データ 2］では参照先情報源のすべての文書に対しキーワードの正解率を求めた．具体的には，文書に含まれる単語列の重要度 S を計算し，その値が最大となる単語列をキーワードとして選択した場合の正解率と，その値がその文書で上位 5%以内の単語列をキーワードとして選択した場合の正解率を求めた．前者は，ハイパーリンク自動生成システムにおいて各文書から重要度が最大となる 1 個の単語列を選びアンカーテキストとする場合に該当し，後者は上位 5%以内の単語列をアンカーテキストとする場合に該当する．

ここで，正解率は，実験対象の 200 文書から従来方式あるいは提案方式で選択したキーワードの集合を A，A のうち確認により文書のトピックと認められたキーワードの集合を B としたとき，|B|/|A|で求める．この正解率の規準では，上位 5%以内をキーワードとして選択したとき，長い文書(含まれる単語列の数が多い文書)の正解率への影響が大きくなる．そのため，本実験では，各文書で重要度の値が上位 5%以内のキーワードが 5 個以上存在する場合は，上位 5 個のみをキーワードとして選択することとした．

ここで，正解の判断基準を説明する．1.1 節で述べたように，キーワード抽出手法として統計的手法を選択したのは，問合せ回答のための情報収集で参照の起点となるキーワード(例．操作名，障害現象名，部品名，エラーコード名)は，文書の中心的な役割(トピック)を果たしている可能性が高いと考えられるためである．したがって，従来手法あるいは提案手法で選択したキーワードが，これらに当てはまる場合は正解と判断した．具体的には，操作名(例.カートリッジ交換)，障害現象名(例．紙詰まり)，部品名(例.センタートレイ)，エラーコード名(例．コール 07)，機能名(例．モノクロ印刷)，設定値名(例.カラーバランス)を正解のキーワードと認定した．

なお，従来手法あるいは提案手法で選択したキーワードが正解のキーワードの部分一致となっているときは，部分一致であっても正解のキーワードと同一と認識できる場合にのみ正解と判定した．例えば，選択したキーワードが「停電電源」で正解のキーワードが「停電電源装置」の場合，部分一致であっても「停電電源」が装置であることは認識できるため同一であり正解と判断した．

4.4 実験結果・考察

表 1 に[実験データ 1]に関して重要度 S が最大となる単語列をキーワードとした場合の正解率と，重要度 S が上位 5%の単語列をキーワードとした場合の正解率を示す．BS の行が従来方式の正解率を，[BS+リンクの特徴 1]と[BS+リンクの特徴 2]の行が提案方式の正解率を表す．また左側の列が「プリンタ A」に関する参照先文書セットの正解率，右側の列が「プリンタ B」の正解率を表す．同様に表 2 に[実験データ 2]に関する実験結果を示す．

提案方式は，全ての結果において従来手法と比較して正解率が 3%から 16%向上している(重要度が最大となる単語列をキーワードとしたときの方が差異は大きい)．このことから，適切な情報源ペアに対して従来方式を補正する提案方式は正解率向上に貢献できるといえる．[実験データ 2]におけるリンクの特徴 2 を用いた重要度計算手法の効果が[実験データ 1]よりも小さいのは，参照先文書の数が少ないことにより重要度の補正項の影響が小さくなったためと考えられる．また，[実験データ 2]の「サーバ B」の従来手法に対する効果がその他(「サーバ A」や[実験データ 1])よりも劣るのは，ベースとする tf/idf の性能が低いためと考えられる．提案方式による補正項が有効に働いた場合でも，ベースとする tf/idf 値が低い単語列を上位にするのは困難である．

一方，実用面を考慮すると，オペレータが自動生成されたリンクに対して肯定的な印象を保持できる程度の正解率が求められる．不正解が多くリンクが有効でないという印象を一度オペレータが持ってしまうと，リンクの存在を無視するようになるからである．仮に 5 回に 1 回までの誤りは許容できると考えると正解率 80%が必要となる．この場合，本実験では，上位 5%を選択したとき正解率が最大でも 66%に留まり，さらなる正解率の向上が必要となる．1.3 節で述べたように，高精度のキーワード抽出を実現するためには，様々な指標を組み合わせる必要があると考える．一つの方法として，本実験ではリンクの特徴 1 と特徴 2 を個別に導入したが，双方を導入することで正解率向上が期待できる．また文献[12]では，技術名や製品名等の企業内情報共有に必要なキーワードを抽出するために，これらのキーワードの種類と，各々の種類のキーワードの内部あるいは周辺に現れやすい表現を抽出ルールとして定義し，抽出ルールに適合する回数を指標として tf/idf と組合せて使用している．本研究では，コンタクトセンターの問合せ回答に必要となるキーワードの種類を定めており(部品名，障害現象名，エラーコード名，操作名等)，抽出ルールを定義すれば同様の指標を導入することは可能と考える．「エラーコード名はコールや Error といった文字列を含む」あるいは「障害現象名の後方には”が発生”といった文字列が続く」のように比較的すぐに思いつく抽出ルールは存在するため，少数の抽出ルールを適用するだけでも，導入/メンテナンス容易性を維持しながら正解率の向上は可能と考える．これらの実用性能達成のための正解率向上は今後の課題とする．

提案方式による重要度の補正が有効であることから，適切な情報源ペアの参照元文書のキーワードには，対応する参照先文書が存在する場合が多いことがわかる．したがって，今回の実験結果は提案手法が前提とするリンクの特徴の一部の性質を検証できたといえる．ただし，その参照先文書が少数であること，そのキーワードが多数の参照元文書に含まれていることが，キーワード抽出の性能向上に貢献するかかは，今回の評価結果からでは明らかではなく，今後より詳細な調査が必要である．

5. おわりに

本研究では，従来のハイパーリンク生成方式における(1)参照先文書の絞込み，(2)キーワード抽出性能の向上の課題の解決を目的として，課題解決型のハイパーリンク生成方式を提案した．また，課題解決型ハイパーリンク生成における適切な情報源ペアに出現するリンクの特徴を用いた新しいキーワード抽出方式を提案した．製品保守の窓口業務で使用する文書を用いて提案方式を評価したところ，適切な情報源ペアに対して従来方式(tf/idf)を補正する提案方式は正解率向

表 1 実験データ 1 の実験結果(*はカイ二乗検定による有意確率 p<0.05, **<0.01)
Tab. 1 Experimental Result of Data Set 1

■重要度Sが最大となる単語列をキーワードとしたときの正解率

方式	プリンタA	プリンタB	総計
BS	56%	51%	53%
	(111/200)	(101/200)	(212/400)
BS+	64%	61%	62%**
リンクの特徴1	(128/200)	(121/200)	(249/400)
BS+	67%	60%	64%**
リンクの特徴2	(134/200)	(120/200)	(254/400)

■重要度Sが上位5%の単語列をキーワードとしたときの正解率

方式	プリンタA	プリンタB	総計
BS	50%	49%	50%
	(322/642)	(333/680)	(655/1322)
BS+	57%*	57%**	57%**
リンクの特徴1	(368/642)	(388/680)	(756/1322)
BS+	57%*	56%*	56%**
リンクの特徴2	(368/642)	(378/680)	(746/1322)

表 2 実験データ 2 の実験結果(*はカイ二乗検定による有意確率 p<0.05, **p<0.01)
Tab. 2 Experimental Result of Data Set 2

■重要度Sが最大となる単語列をキーワードとしたときの正解率

方式	サーバA	サーバB	総計
BS	62%	41%	53%
	(39/63)	(19/46)	(58/109)
BS+	78%	52%	67%*
リンクの特徴1	(49/63)	(24/46)	(73/109)
BS+	75%	50%	64%
リンクの特徴2	(47/63)	(23/46)	(70/109)

■重要度Sが上位5%の単語列をキーワードとしたときの正解率

方式	サーバA	サーバB	総計
BS	60%	41%	52%
	(166/276)	(79/195)	(245/471)
BS+	66%	44%	57%
リンクの特徴1	(182/276)	(85/195)	(267/471)
BS+	63%	44%	55%
リンクの特徴2	(174/276)	(85/195)	(259/471)

上に貢献することが判明した.

[謝辞]

本研究を進めるにあたり，評価用のデータを提供，及び，応対記録から関連文書へのハイパーリンクが窓口業務に有効である旨の助言を頂いたNECフィールディング株式会社東日本カスタマサポート本部殿，及び，NEC共通基盤ソフトウェア研究所の中川淳子様に深く感謝致します.

[文献]

[1] 黒橋 禎男, 長尾 真, 佐藤 理史, 村上 雅彦, 専門用語辞典の自動的ハイパーテキスト化の方法, 人工知能学会誌, Vol.7, No.2, pp.336-345, 1991.
[2] 産経 MSN ニュース, http://sankei.jp.msn.com/
[3] 石田 和生, 市山 俊治, 複数文書間のハイパーリンク自動生成とメンテナンス. 情報処理学会研究報告 デジタル・ドキュメント 17-5, pp.33-40, 1999.
[4] 大森 信行, 岡村 潤, 森 辰則, 中川 裕志, 情報検索手法を利用した関連マニュアル群のハイパーテキスト化, 情報処理学会論文誌, Vol.40, No.6, pp. 2776-2784, 1999.
[5] 服部 元, 原 隆浩, 滝嶋 弘康, 菅谷 史昭, 西尾 章治郎, 周辺語を活用したクリック型 Web 検索システムの提案と評価, 情報処理学会論文誌 データベース, Vol.1, No.2, pp. 26-37, 2008.
[6] 関根 聡, 固有表現から専門用語, 言語処理学会第 10 回年次大会 (NLP2004)「固有表現と専門用語」 ワークショップ, 2004.
[7] Sparck-Jones, K., A Statistical Interpretation of Term Specificity and Its Application in Retrieval, Journal of Documentation, 28(1), pp.11-21, 1972.
[8] 中川 裕志, 湯本 紘彰, 森 辰則, 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, Vol.10, No.1, pp. 27-46, 2003.
[9] Frantzi, K. and Ananiadou, S., Extracting Nested Collocations, Proceedings of the 16th International Conference on Computational Linguistics (COLING 96)", pp. 41-46, 1996.
[10] 久光 徹, 丹羽 芳樹 and 辻井 潤一, タームの representativeness を測る, 情報処理学会研究報告 自然言語処理研究会, NL-133-16, pp. 115-122, 1999.
[11] 形態素解析システム茶筌, http://chasen.naist.jp/hiki/ChaSen/
[12] 立石健二, 久寿居大, 複数の作成者情報付き文書からの専門用語抽出, 情報処理学会論文誌:データベース, Vol.47, No.30, pp.24-32, 2006.

立石 健二 Kenji TATEISHI
1999 年九州大学大学院システム情報科学研究科知能システム学専攻修士課程修了，同年 NEC 入社．現在，NEC 共通基盤ソフトウェア研究所主任．自然言語処理の応用システムに関する研究に従事．情報処理学会，言語処理学会，日本データベース学会，各会員.

細見 格 Itaru HOSOMI
1993年神戸大学大学院工学研究科システム工学専攻修士課程修了，同年NEC入社．現在，NEC共通基盤ソフトウェア研究所主任研究員．自然言語処理，セマンティックウェブ，オントロジーに関する研究開発に従事．情報処理学会会員.

久寿居 大 Dai KUSUI
1992 年京都大学大学院工学研究科修士課程修了. 現在, NEC 共通基盤ソフトウェア研究所主任研究員．情報分析・知識管理システムの研究・開発に従事．情報処理学会会員.