ビッグデータ活用のためのテキスト分析技術

土田 正明・石川 開・久寿居 大
楠村 幸貴・中尾 敏康

要　旨

ビッグデータに含まれる大量のテキストは、人が人に情報や意図を伝えるために作成されたデータであるため、価値の高い情報が含まれる重要な情報源です。NECでは、大量のテキストデータから「顧客の声」や「風評」を抽出して、マーケティング、企業リスク管理、顧客管理などに活用するための技術開発に取り組んでいます。本稿では、最近の研究開発成果から文間の意味の包含関係を認識するテキスト含意認識技術、サイバー情報からの風評検知技術、コンタクトセンター業務を効率化する意味検索技術を紹介します。

キーワード

●テキスト分析　●意味検索　●風評検知　●テキスト含意認識

1. まえがき

テキストは、人が人に情報や意図を伝えるために作成されたデータです。そのため、新聞、雑誌、Webページ、社内文書、電子メールなど社会に存在する大量のテキストデータは、ビッグデータの中でも、人にとって価値の高い情報を含む情報源として重要です。一方で、大量のテキストデータを活用するためには、a）表現の曖昧性を解消しながら、b）含まれている意味や他の情報との関係を理解したうえで、c）大量のデータから必要なものを的確かつ迅速に見つけて加工する、という技術が必要となります。NECでは、商品・サービス・人物などに関する「顧客の声」や「風評」を抽出して、マーケティング、企業リスク管理、顧客管理などを実現することを目的に、次の研究開発を進めています（**図1**）。

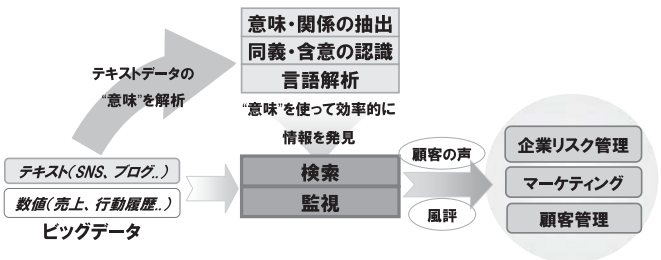

図1　大量テキストから価値を抽出する技術

(1) 話し言葉など自由な表現に対応する言語解析技術
(2) 単語や文が同じ意味かどうかを判定する同義・含意判定技術
(3) 意味や関係を抽出する意味・関係抽出技術
(4) 抽出した意味に基づき多様な情報源を統合活用する検索技術や監視技術

本稿では、最近の研究成果から、(2) に関係する文間の意味の包含関係を認識するテキスト含意認識技術、(4) に関係する、サイバー情報からの風評情報検知技術、コンタクトセンター業務を効率化する意味検索技術、を紹介します。

2. 意味的包含関係を認識するテキスト含意認識技術

テキストデータでは、同じ意味の内容が異なる表現で書かれている場合がたくさんあります。異なる表現が同じ意味であることを認識する技術は、テキスト含意認識技術（Recognizing Textual Entailment：RTE）と呼ばれます。弊社のRTEは、2011年に米国国立標準技術研究所（National Institute of Standards and Technology：NIST）が主催した評価ワークショップTAC2011 RTE-7で第1位を獲得しました。

RTEは、2つのテキストを入力として、片方のテキストがもう片方のテキストを意味的に含む場合に、含意関係と認識する技術です。例えば「A社の社長がニューヨークに出張した」というテキストは、「A社の社長がアメリカに行った」という

内容を意味的に含んでいます。これは「ニューヨークに出張した」ということは、確実に「アメリカに行った」ことになるためです。

RTEによって、特定の意味を含むテキストを検索したり、逆に特定の意味を含むテキストがないことで、その意味の内容を「新規情報」として検出したり、同じ意味のテキストをまとめるといったことが可能になります。

弊社の技術では、文中の単語の重要性、主語や述語などの文の構造を考慮することで、高精度な含意関係の認識を実現しています。具体的には、2つのテキストにおける単語間の表現の違いを考慮しておおまかに意味が一致することを判定し、次に、文の構造から実際には含意関係にはないものをふるい落とすという、2段階の判定処理を行っています。これによって、異なる単語で同じ意味が表現されている場合と、同じ単語で異なる意味が表現されている場合に対応し、認識の誤りを防いでいます。

TAC2011 RTE-7では、与えられた2つのテキストが含意関係にあるか否かを判定するメインタスクと、あるテキストが与えられたテキストデータ中に存在しない、すなわち、新規情報であるか否かを判定するサブタスクの両方で1位を獲得しました。

現在、弊社ではRTEをマーケティングや企業のリスク管理に活用すべく研究開発を進めています。また、これまでは顧客の声の傾向を調べるために特徴的なキーワードを分析したり、風評情報を検知するためにキーワードを設定したりと、人間がコンピュータに合わせる不自然な状態でしたが、RTEは、人間にとって自然な文のままコンピュータ上でテキストを扱うための基本技術となります。

3. サイバー情報からの風評検知技術

ブログやソーシャルネットワーキングサービスに代表される、インターネット上での情報発信やコミュニケーションの手段が普及したことによって、噂や悪評が素早く広まることで損失を被る、いわゆる風評被害が顕在化しています。例えば、ある銀行は、破綻の噂がチェーンメールで拡散した結果、数百億の預金が引き出される被害にあっています。こういった風評被害を食い止めるには、風評になり得るリスク情報を早期に検知することが重要です。

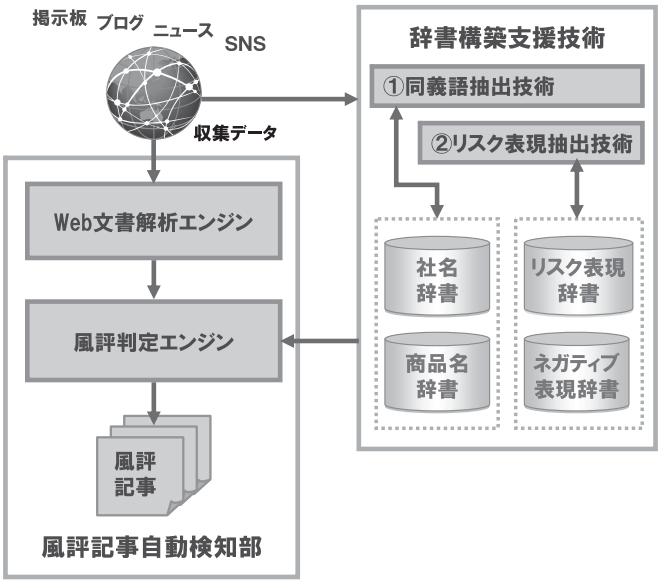

図2　風評検知システムの構成

弊社ではインターネット上の情報（以下、サイバー情報）から風評のリスク情報を検知するシステムを開発しました（**図2**）。風評検知システムは、風評検知のための辞書の作成を支援する「辞書構築支援技術」と、大量のテキストデータから辞書に登録されている表現を含む記事を風評リスク情報として検知する「風評記事自動検知部」から構成されています。辞書は「社名辞書」と「商品名辞書」が風評の対象に、「リスク表現辞書」と「ネガティブ表現辞書」が風評の内容に、それぞれ対応します。例えば、「社名辞書」に「A銀行」、「リスク表現辞書」に「破たん」という表現が登録されている場合、大量のテキストデータ中から「A銀行が破たんの噂」や「A銀行は不良債権が原因で破たんするらしい」などのテキストを検知できます。

リスク情報を漏れなく検知するためには、検知したいリスクの表現と、そのバリエーションを辞書に登録することが重要です。このような辞書を人手で作成することは、大量のテキストを読んで該当する表現を抜き出すなど、大変な労力が必要となるため、現実的ではありません。

辞書構築支援技術は、少ない労力で辞書の作成を行えることを目的に開発したもので、「同義語抽出技術」と「リスク表現抽出技術」から構成されます。

同義語抽出技術では、同じ意味の単語を抽出することで、表現のバリエーションを増やすことができます。例えば、「A銀行」と「破たん」という表現だけでは、「A銀が破綻するらしい」は検知できませんが「A銀行」と「破たん」の同義語である「A銀」、「破綻」が登録されていれば検知できます。開発技術は、こういった同義語を、省略語らしさ（「A銀行」と「A銀」）、翻訳語の一致（「破綻」と「破たん」の訳語は共に“bankruptcy”）、表記の類似性（「外貨預金」と「外貨貯金」）など、さまざまな基準で抽出します。また、翻訳して省略語を生成（外貨預金→“foreign currency deposit”→“FCD”から「外貨預金」と「FCD」）するなど、多段階変換も行います。これらによって、文字列としては似ていない同義語も含めて、多くの候補を抽出できます。

リスク表現抽出技術は、リスクを表す表現を抽出することで、リスク表現辞書の作成を支援します。例えば、銀行のリスクは「破たん」だけではなく「架空請求」「フィッシング」なども考えられます。開発技術は、同業他社などの風評リスク情報が含まれる記事と一般的な記事を入力として、それぞれに含まれる表現とその出現数を計算し、その統計的な差からリスク記事に現れやすい表現を抽出します。

また、開発技術ではリスク表現を適切な順にランキングします。例えば「フィッシング」に関するリスク表現を抽出する場合、「メール」という表現はリスク記事以外でも多く出現するためリスク表現としては不適切ですが、「不審なメール」はリスク記事に特に多く出現するため、リスク表現として適切と考えられます。一方、「不審なメールが送付」は、リスク表現として適切であるものの、「不審なメールが銀行から」などのテキストを検知できなくなってしまう問題があります。開発技術では、「メール」はリスク表現としては不適切な表現として除外し、「不審なメール」と「不審なメールが送付」では、「不審なメール」をより上位にランキングして出力します。

辞書の作成者は、辞書構築支援技術より出力される表現を確認し、望ましい表現を辞書に追加していくだけで、辞書を作成できます。表現の確認は、テキストを読む作業と比べて少ない労力で行えるため、辞書作成の人的コストを軽減できます。評価実験では、テキストを読みながら人手で作成した場合と同等の検知精度の辞書を、約60%の時間で作成できました。また、開発した風評検知システムによって、約6,000語の辞書とサーバ1台を用いた場合、約1,800万のTwitter上のテキスト（以下、ツイート）を約2.5時間で処理でき、企業名が含まれるツイートを人がすべて読む場合と比べ、コストを約30%削減できました。

4. コンタクトセンター業務を効率化する意味検索技術

コンタクトセンターに蓄積されている膨大な問合せ応対記録を活用し、正確で迅速な回答を可能にする検索技術を開発しました（**図3**）。

一般に、技術サポート系コンタクトセンターでは、オペレータがお客様からの問合せ内容を把握し、大量に蓄積されたナレッジ文書（過去事例、マニュアル、技術情報など）を参照して、原因や解決策を特定する必要があります。しかしながら、サポート対象製品の拡大や製品機能の多様化により、お客様の問合せ内容が複雑かつ高度化しており、情報の分析や的確な情報の取得に時間を要し、お客様への回答に時間が掛かるケースがあります。

意味検索技術は、検索キーワードに関連するテキストを網羅的に検索することで検索漏れを少なくする技術で、具体的には検索する言葉をその同義語や上位・下位概念の言葉にも広げて、関係する文書を幅広く検索します。例えば、検索キーワード（例“OS”）に対して、同義語（“オペレーティングシステム”）、上位概念（“ソフトウェア”）、下位概念（“Linux”）など意味的に関係のあるキーワードを含む文

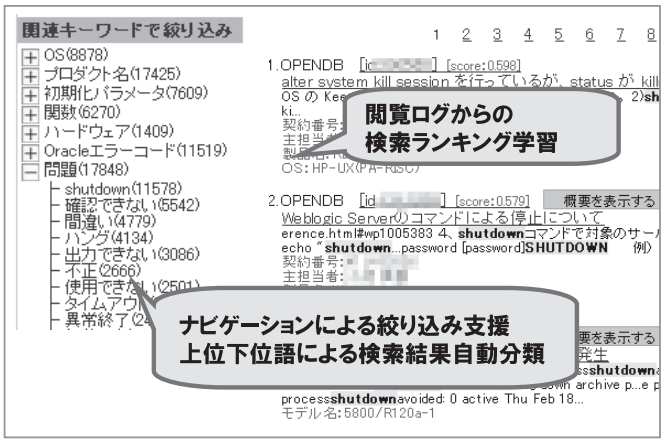

図3　ナレッジ検索システムの画面例（部分）

書も同時に検索できます。弊社の意味検索技術の特徴は次のとおりです。

- **意味検索を高速に実行可能な圧縮インデックス管理**
  上位・下位概念を含んだ検索キーワードと文書との関係を記載したインデックスのデータを圧縮保持しています。コンパクトなインデックスにより、検索キーワードを上位・下位概念を使って展開する場合に、オンメモリ処理が可能となり、従来と比べて圧倒的に高速な検索処理が実現できます。
- **ナビゲーションによる絞り込み支援方式**
  あらかじめ用意された大規模な上位下位概念辞書に基づき、各検索結果に含まれる重要語を用いて検索結果を自動分類します。これにより、オペレータの知識には依存せずに絞り込みのためのキーワード候補が概観できるようになるため、初級クラスのオペレータでも検索結果を容易に絞り込めるようになります。
- **閲覧ログからの検索ランキング学習方式**
  サーバ上に記録された検索及び閲覧ログ（検索キーワードと閲覧された文書を組み合わせたデータ）と、オペレータが入力した各文書に対する評価データとを組み合わせて、検索キーワードと文書の有用度の関係を学習します。これにより検索精度が改善されます。

これらの技術を組み込んだ検索システムを、技術サポート系コンタクトセンターであるNECオラクルレスポンスセンターに適用して、効果を評価しました。その結果、導入前後を比較すると、問合せ件数が増加しているにも関わらず、平均的な対応完了までの日数（以下、平均TAT）は減少し、お客様から評価いただく顧客満足度も向上していることが確認できました。

具体的には、問合せ件数は31%増加しましたが、平均TATは19.1%減少し、顧客満足度は「満点」の評価が7.8%増加していました。また、オペレータの作業時間でみると、中級クラスでは1回あたりの検索時間が平均14％減少し、かつ、1日あたりの検索回数は28%増加しました。初級クラスでは1回あたりの平均検索時間は3%の減少にとどまりましたが、1日あたりの検索回数は95%増加しました。

これらの結果より、検索時間が減少し、検索回数が増加することで迅速かつ丁寧な応対が実現され、平均TATの短縮と顧客満足度の向上を同時にもたらしたと考えられます。

5. むすび

本稿では、大量のテキストデータを活用するための自然言語処理技術として最近の研究開発成果を紹介しました。ビッグデータの活用にはテキスト、数値などの非テキストの双方を合わせて分析する技術が重要になると考えられます。特に、さまざまな意志決定を行う場面では分析の結果だけではなく、結果に至る理由や結果そのものの解釈を支援する情報も必要となります。テキストはこのような情報を提供するために重要な情報源であり、弊社では引き続き、関連する技術の研究開発に取り組みます。

執筆者プロフィール

土田 正明
中央研究所
情報・ナレッジ研究所
主任

石川 開
中央研究所
情報・ナレッジ研究所
主任研究員

久寿居 大
中央研究所
情報・ナレッジ研究所
主任研究員

楠村 幸貴
中央研究所
情報・ナレッジ研究所
主任

中尾 敏康
中央研究所
情報・ナレッジ研究所
研究部長

関連URL

NEC、文書の高速検索と分類表示を行う意味検索エンジンを開発:
http://www.nec.co.jp/press/ja/1202/2702.html
NECのテキスト含意認識技術が米国国立標準技術研究所（NIST）主催の評価タスクにおいて第一位を獲得:
http://www.nec.co.jp/press/ja/1204/1301.html

NEC 技報のご案内

NEC技報の論文をご覧いただきありがとうございます。
ご興味がありましたら、関連する他の論文もご一読ください。

NEC技報WEBサイトはこちら

NEC技報(日本語)　　NEC Technical Journal(英語)

Vol.65 No.2　ビッグデータ活用を支える基盤技術・ソリューション特集

ビッグデータ活用を支える基盤技術・ソリューション特集によせて
ビッグデータを価値に変えるNECのITインフラ

◇ 特集論文

データ管理 / 処理基盤

超高速データ分析プラットフォーム「InfoFrame DWH Appliance」
SDN技術で通信フローを制御する「UNIVERGE PFシリーズ」
大量データをリアルタイムに処理する「InfoFrame Table Access Method」
大量データを高速に処理する「InfoFrame DataBooster」
ビッグデータの活用に最適なスケールアウト型新データベース「InfoFrame Relational Store」
高い信頼性と拡張性を実現したExpress5800/スケーラブルHAサーバ
大規模データ処理に対するOSS Hadoopの活用
大容量・高信頼グリッドストレージiStorage HSシリーズ(HYDRAstor)

データ分析基盤

ファイルサーバのデータ整理・活用を支援する「Information Assessment System」
超大規模バイオメトリック認証システムとその実現
WebSAMの分析技術と応用例～インバリアント分析の特長と適用領域～

データ収集基盤

スマートな社会を実現するM2Mとビッグデータ
微小な振動を検知する超高感度振動センサ技術開発とその応用

ビッグデータ処理を支える先進技術

多次元範囲検索を可能とするキーバリューストア「MD-HBase」
高倍率・高精細を実現する事例ベースの学習型超解像方式
ビッグデータ活用のためのテキスト分析技術
ビッグデータ時代の最先端データマイニング
ジオタグ付きデータをクラウドでスケーラブルに処理するジオフェンシングシステム
柔軟性と高性能を備えたビッグデータ・ストリーム分析プラットフォーム「Blockmon」とその使用事例

◇ 普通論文

地デジTVを活用した「まちづくりコミュニティ形成支援システム」

◇ NEC Information

NEWS

スケールアウト型新データベース「InfoFrame Relational Store」が2つの賞を受賞

Vol.65 No.2
(2012年9月)

特集TOP