自動記事分類技術を用いた「日経テレコンナビ型記事検索」 情報収集に気づきを提供する新スタイル

New Nikkei Telecom navi type style for article search with automatic classification techniques

出口 信吾¹

DEGUCHI Shingo¹

.....

1 株式会社日本経済新聞デジタルメディア テレコン事業本部営業部 (〒100-8066 東京都千代田区大手町1-3-7)

Tel: 03-6256-2570 Fax: 03-6256-2548 E-mail: shingo.deguchi@nex.nikkei.co.jp

1 NIKKEI DIGITAL MEDIA INC. Sales & Marketing Department NIKKEI TELECOM DIVISION (1-3-7 Otemachi Chiyoda-ku, Tokyo 100-8066)

原稿受理 (2014-12-26)

情報管理 57(12), 900-909, doi: 10.1241/johokanri.57.900 (<http://dx.doi.org/10.1241/johokanri.57.900>)

著者抄録

インターネット経由でビジネス情報提供を手がける日経テレコンはサービス開始から 30 周年を迎えた節目の 2014 年 4 月に、自動記事分類技術という新しい技術を導入したナビゲーション型の記事検索インターフェース「ナビ型記事検索」をリリースした。従来のキーワードによる全文検索に比べて、検索結果の適合率を大幅に改善すると同時に、利用者が自分の意図を反映しながら検索結果を絞り込んでいくことができる、新しい検索スタイルを提案するものである。本稿では、「ナビ型記事検索」を使った検索手法を紹介しながら、それを支えている自動記事分類技術の概要と検索用インデックスの生成処理についても解説する。

キーワード

日経テレコン, 記事, 検索, 自動分類, ナビゲーション型, 言語解析, 言語理解, キーワード, シソーラス, インデックス

1. はじめに

2014年4月、日経テレコン (<http://t21.nikkei.co.jp>, 日本経済新聞デジタルメディア) はサービス開始から30周年を迎えた。この間、「信頼のおけるビジネス

情報を提供し、経済活動に貢献する」という使命のもと、有用なコンテンツの収録を拡大しながら、必要な情報を利用者が探しやすい検索技術を追求してきた。そして30周年の節目に「ナビ型記事検索」という新しいメニューをリリースした。新メニューで

は、検索にヒットした記事を自動的に分類して利用者の意図に沿った情報を提供するインターフェースを実現している。この自動記事分類の技術を、従来の「インデックス作成」や「全文検索技術」と組み合わせた仕組みについて本稿で紹介する。

1.1 日経テレコンとは

日経テレコンはインターネットで利用できる有料の会員制ビジネス情報サービスである。主力メニューの1つ「記事検索」では、505種類（2014年12月現在）に及ぶメディアに掲載された記事をまとめて検索することができる。全国紙・業界専門紙（誌）に加えて、2014年末には全都道府県の50紙をカバーした。日本語で読める海外情報にも注力しており、アジアビジネスに関しては30種類のメディア情報を収録している。また「企業検索」では、国内外の主要な調査会社による企業・財務情報などをワンストップで収集できる。「人事検索」で提供する「日経WHO'S WHO」は国内ビジネスパーソン約30万人の情報が検索できる。

1.2 生まれたときは「日経テレコム」

日経テレコンの誕生は1984年4月2日で、当初は「日経テレコム」という呼称であった¹⁾。まだオフィスでもパソコンが物珍しかった当時、専用端末でサービス提供を開始した。その後にはパソコンからモデム通信で利用できる情報サービスとして注目された。

そして、インターネットがビジネスに普及しはじめた1997年、Web時代に即応した日経テレコン21の提供を開始した。インターネットの活用によりビジネス現場からアクセスできる情報量が爆発的に増え続けている。特に広告掲載をビジネスモデルとしているWebサイトを通じて提供されるニュースをはじめ、企業のWebサイトや個人ブログまで、発信される情報には誰でもアクセスできるようになっている。

一方で、ビジネス現場における情報収集では、内容の正確性と情報の鮮度が求められる。効率よく必要な情報を探し出し、次のビジネスアクションに取り

掛かるためには、業務目的に合ったコンテンツが収録されているデータベースを使うことが有効である。

日経テレコンの記事検索の特徴は、新聞や雑誌など信頼性の高いメディアに掲載された情報だけを収録してデータベースを構築していることである。注目のトピックス、企業、人物や地域の名称など、ビジネスシーンに合った自由なキーワードでの確な情報収集を可能にしている^{2), 3)}。

2. “新しい検索スタイル”を提案する「ナビ型記事検索」のリリース

2014年4月に、新しい記事検索メニュー「ナビ型記事検索」の提供を開始した（図1）。

利用者が思いついたキーワードを起点として、検索された結果を複数の観点で自動的に分類し、利用者が求めている情報（検索結果）にたどり着くことができるようになっていく。「ナビ型記事検索」の「ナビ」は、利用者を適切な検索結果にナビゲーションすることを意味したもので、初めてでも無理なく検索できるよう、新たな検索スタイルを提供することをコンセプトにしている。

2.1 従来型の記事検索の課題

日経テレコンの記事検索は、新聞・雑誌などメデ

図1 「ナビ型記事検索」の検索画面

アに掲載された記事を中心に収録し、2014年12月時点で505種類の媒体をカバーしている。データベースには最新の記事情報が日々追加されるうえ、収録する媒体数も増え続けており、2014年11月7日には、収録記事の本数がついに1億本を超えた。

新聞・雑誌や海外ニュース、業界専門情報などのアーカイブを対象に、キーワード1つで横断検索できるメリットは非常に大きい。一方で、検索結果として表示される記事数が多くなり、本当に欲しい情報にたどり着くまでに時間がかかる傾向が出てきている。

また、日経テレコンの草創期は情報収集のスペシャリスト（サーチャー）が記事検索を行っていたが、インターネット時代になり、日経テレコンもビジネスの現場で営業担当者や与信管理担当者が直接利用するケースが増えてきた。それに伴い、検索スキルを備えていない利用者が思いついた言葉だけで検索を進めても目的が達成できない、という場面が多々みられるようになっている。

2.2「ナビ型記事検索」の狙いと特徴

「ナビ型記事検索」と従来型の記事検索との違いは、一言でいうと検索結果の見える化である。「キーワードを入れても欲しい情報が見つからない」「たくさんの検索結果から欲しい情報だけ取り出したい」という課題を解決するために、「ナビ型記事検索」では、思いついたキーワードから探したい概念に絞り込めたり、検索結果を分類したりする言語理解的なアプローチを取り入れている。

さらに、利用者の視点に立って、使いやすさとわかりやすさを重視して、思い切ってシンプルな検索画面にしている。「ナビ型記事検索」のメニューを選ぶと、キーワードの入力ボックスのほかは、期間選択バーなど必要最小限の設定パーツのみを表示している。これまでの記事検索画面を見慣れた利用者には物足りなさを感じるかもしれない。提供する機能を画面上で漏れなく提示する方がサービスを提供す

る側としては安心だが、利用者にとっては必要なものと不要なものの区別がつかず、かえって混乱することもあったはずと考えた。

反対に、検索結果については情報量を増やしている。「何件の記事が見つかった」という情報だけでなく、検索結果をさまざまな軸で分類したり、時間軸で検索結果の分布をグラフ表示したりしている。利用者が考えたキーワードだけで検索すると「どうしてこの記事がヒットしたのだろうか」と疑問に思うこともあったはず。検索結果を分類や時間軸で仕分けしておくことで、意図する内容にさらに絞り込むようナビゲートしている。

以下に「ナビ型記事検索」の特徴をまとめた。

(1) キーワードの属性／種別を判定するサジェスト機能

従来のように単に文字列が一致する言葉を提示するのではなく（図2）、その言葉の属性（会社、団体、人物、一般など）も併せて表示（図3の囲み枠）して、よりの確なキーワードを選定できるようにした。

(2) 検索された記事群に含まれる主題語と分類語（主題と分類語については「3.1 言語理解によるタグ付け」を参照）を一覧表示する「絞り込みキーワード候補」機能（図4）。

検索結果の件数が多いと感じるときは、どのような言葉で検索を重ねると、どの程度まで件数が絞り込めるのかが一覧表示されている。入力したキーワードがどのようなテーマで報じられていることが多いのか、一目でわかるよう分類別に件数の多い順に並

図2 従来の記事検索のサジェスト機能の例

入力された言葉が一致する候補を提示するが、その属性までは判別できなかった。

べている。

(3) 検索結果の時系列表示による可視化 (図5)

キーワードの注目度の変遷を追ったり、新語の登場時期を探ったりする、といった用途も考えられる。

図3 サジェスト機能の例

入力された「ライオン」で会社のライオンを検索したいのか、一般用語の「ライオン」を検索したいのか、意図を反映できる。

図4 絞り込みキーワード候補の例

キーワード「LED」で検索し3,494件の記事が見つかったケース（検索対象の媒体、期間の選択により件数は異なる）。3,494件の記事がどのような「主題語」「分類語」で何件に絞り込めるかを提示している。

2.3 絞り込みキーワード候補から得られる新たな気づき

特に前節の(2)で触れた絞り込みキーワード候補を表示する機能は、多すぎる検索結果を最適に絞り込むのに有用であるだけでなく、一覧された候補語を眺めることで、これまで自分が気づいていなかったり、知らなかった用語、概念やトレンドを発見したりすることがある。

たとえば、「LED」で検索したときの絞り込みキーワード候補から、思わぬ用語との組み合わせが浮かび上がってくるかもしれない(図6)。

3. 言語解析から言語理解へ

「ナビ型記事検索」のインターフェースで重要な役割を担っているのが、「分類タグ」(以下、タグ)で、自動記事分類技術を使って各記事に付与される。

3.1 言語理解によるタグ付け

タグには、主題語(表1)と呼ぶ「会社名」「団体名」「人物名」「一般用語」と分類語(表2)と呼ぶ「テーマ」「業界」「地域」「記事種別」のカテゴリーを用意している。

各カテゴリーにおいて最適なタグが付与されるよう、過去の大量の記事を解析した結果を反映した知識ベースを構築し、言語理解的アプローチによる検索用インデックスの生成処理を実現している(図7)。

図5 検索結果の時系列表示の例

検索された3,494件の記事を時系列に仕分けたグラフを表示。グラフをクリックすると、検索結果を当該期間のものだけに絞り込むことができる。また対象期間を、グラフ下のスライダー(囲み)をマウスで操作して設定することも可能である。

図6 キーワード「一般：LED」での検索例

タグの横の数字は、そのタグで絞り込んだときの検索結果の件数を表す。

表1 主題語

会社名	上場会社、非上場（日経会社コードつき）会社、その他主要会社
団体名	省庁、大学、政党などの団体名
人物名	日経WHO'S WHOや記事頻出の著名人
一般用語	その他名詞

表2 分類語

テーマ	記事の主題を表す内容別の分類。分類体系は「企業活動」（大分類：企業）と「企業を取り巻く環境」（大分類：政治・経済・技術・社会）から構成。
業界	記事の主題と関連する業界別の分類。日経新業種分類をもとに定義。
地域	記事の主題と関連する地域別に「海外地域」「国名」「国内地域」単位で分類。
記事種別	記事のタイプ別に7種類に分類。インタビュー、調査・統計、社説…など

記事ごとに付与されたタグは、記事本文を表示した画面で確認できる。「その他の書誌事項を表示」をクリックする（図8のa）と、カテゴリー別に生成されたタグが表示される。

たとえば、2014年7月9日付日本経済新聞朝刊の記事「東大・日経の物価指数、脱デフレ、検証の一助に、動向きめ細かく把握。」では、キーワード抽出処理から「会社名」「団体名」「一般用語」のタグが付与され、自動記事分類技術から「テーマ」「記事種別」のタグ

図7 検索用インデックスの作成

インデックス抽出処理と自動記事分類を組み合わせて、多様なタグを検索用インデックスとして付与する。

が付与されている（図8のb）。

これまでの成果として、各カテゴリーのタグは90%以上の精度で適切に付与されていることがわかった。中でももっとも精度が求められるのが、日経会社コードをもつ会社（約2万社）が記事中に主題

図8 タグの付与例

記事本文を表示する画面にて記事ごとに付与されているタグを確認できる。

語として登場したときに、正しくそのコードをタグとして付与することである。同じ会社名が複数存在する場合は、記事本文中に登場する住所や社長の氏名などを手掛かりに判定することもあり、会社名の同定には、同様の精度を実現している。

3.2 検索用インデックス採否プロセスの変化

言語解析処理では、形態素解析^{注1)}をいかに正しく行うか、そしていかに多くの言葉をカバーした辞書とシソーラスを整備しておくか、という点が重要である。

ただし、この手法では辞書に未登録の言葉はインデックスとして採用されない可能性が高い。また同表記異義語がある場合はどのシソーラス語を適用すべきか、という判断を機械的に行うことは困難で、人手による正誤チェックが欠かせない。そして、人手作業によるメンテナンスを待っていると、正しい用語を検索サービスに反映させるまでに時間を要してしまう課題も抱えていた。

一方、今回新たに構築した言語解析処理の最大の特徴は、知識ベースによるルールにもとづいて、辞書に登録されていない用語を認識して抽出できる点

にある。辞書はそれが可能となるように、単語を限定し構造化して作られており、むやみに単語を登録するような作業は行っていない(図9)。

このルールによる検索用インデックスの識別のもっともわかりやすい例が「人名(フルネーム)」の抽出処理である。

辞書にフルネームで登録されている人名は12万語である。しかし、インデックス生成処理では、姓と名の情報からフルネームが動的に検索用インデックスとして生成可能になっている。

たとえば、山田という「姓」と亜衣、亜唯、阿衣という「名」により、以下のようなテキストから、人名を正しく抽出して返すことができる。

「山田亜衣さんが…」→人名「山田亜衣」

「山田亜唯さんが…」→人名「山田亜唯」

「山田阿衣さんが…」→人名「山田阿衣」

このとき辞書に「山田亜衣」「山田亜唯」「山田阿衣」が登録されていなくても抽出できるのである。

つまり、フルネームで辞書に登録されている12万語に加え、「姓」として登録されている語と「名」として登録されている語の組み合わせで生成されるすべての人名が検索用インデックスとして抽出可能と

図9 言語解析と言語理解

言語理解的な処理工程を追加して、検索用インデックスの抽出・生成と属性判定を行う。

なっている。

3.3 自動記事分類技術がもたらしたメンテナンス作業への効果

これまでのインデックス生成処理は、形態素解析→品詞判定（構文解析）→辞書マッチングという流れを機械的に行い、その後に入手作業で辞書に未登録語を追加したり、適切なキーワードを付与したりするメンテナンスを行っていた。

これは、同じ概念の事柄であれば決まったキーワードを付与する「統制語」の考え方にもとづいたものであり、統制語は「シソーラス」で単語の包含関係を管理している^{4),5)}。シソーラスはキーワード検索では有用であったが、記事テキストの全文を対象とした検索が一般的となった今日では、利用者への浸透度や語彙の新鮮さに課題があった。

また、検索用インデックスとして採用された統制語が誤っていた場合や、辞書に未登録のため統制語として採用されなかった新登場の単語があった場合は、入手で検索用インデックスを編集したり、辞書へ単語登録したりする作業を行っていた。

ただし、入手作業が及ぶのは、日経各紙（日経朝夕刊、日経産業、日経MJ、日経ヴェリタスなど）までとなり、外部から提供いただいた記事データについては、提供会社から指定された単語・分類語と、機械処理で追加された統制語を採用するにとどまっていた。

新しいインデックス生成処理は、抽出された言葉が辞書に未登録であっても、自動記事分類技術を使った知識ベースでインデックス判定ができるので、日経各紙に限らず、日経テレコンに収録するすべての新聞・雑誌・ニュース媒体を対象に適用することができる。

自動記事分類技術は記事文脈にもとづいて約500（分類語〔テーマ、業界、地域、記事種別〕の合計数）種類程度のタグを付与しているので、記事中に表れていない用語でも文意での検索用インデックスとし

て採用することができる。なおインデックスには分類名と一緒に分類語コードも付与されている^{注2)}。

「テーマ」には企業、経済、政治、社会などにわたって約140のタグがある。過去の記事データを自動記事分類で分析した結果をもとに採用するテーマを定めている（表3）。「業界」では日経NEEDS業種・中分類に基づく63業界をタグとして付与している（表4）。

2014年12月には、この新しい検索用インデックス生成処理法を、記事検索メニューに収録する全媒体に適用するとともに、過去蓄積分についても2008年分まで遡^{さかのぼ}って再インデックス化を行った。入手のメンテナンス作業が従来に比べて省力化される効果もあった。

表3 「テーマ」の例

分類名	分類語コード
事業組み替え	#W10101
共同出資会社設立	#W10102
新規事業進出	#W10103
事業・企業の買収	#W10106
経営統合	#W10107
分社化・グループ再編	#W10108
既存事業の強化	#W10109
既存事業の縮小・撤退	#W10110
事業・企業売却	#W10111
経営破たん	#W10112
資本参加	#W10113
業務提携	#W10114
提携解消	#W10115
会社設立	#W10116
海外進出	#W10117

※「企業＞事業組み替え」分野のタグの一部。

表4 「業界」の例

提供分類	分類語コード
石油・鉱業・エネルギー	#B0010
電力・ガス	#B0020
繊維・紙・パルプ	#B0030
化学	#B0040
鉄鋼	#B0050
非鉄金属・金属製品	#B0060
重機・建機・プラント	#B0070
産業用ロボット・工作機械	#B0080
医療用機器	#B0090
コンピューター・通信機器	#B0110
複写機・プリンター	#B0120
電機・家電	#B0130

4. 「ナビ型記事検索」の評価と課題

「ナビ型記事検索」は、日経テレコンを使い始めたばかりの方には大変好評である。

2014年10月に日経テレコン利用者を対象にアンケートを行った際に、「ナビ型記事検索」利用者の6割近くから「使い始めから戸惑うことはほとんどなかった」という回答を得た。また、「ナビ型記事検索」の魅力点として、

- ・ キーワードを入力すると候補語が表示される。
- ・ 検索した後に「絞り込みキーワード候補語」が表示される。
- ・ 検索結果の件数が対象期間でグラフ表示される。

の順で多くあげられていた。

その一方で、従来型の記事検索を使い慣れている利用者は、なかなか「ナビ型記事検索」の方に定着しづらいようで、次のような意見を得た。

- ・ 慣用的に使用されている言葉をキーワードにすると、検索が難しい場合がある。
- ・ 使いたいと思う候補語が表示されていない。
- ・ 検索結果から不必要なものを“除く”ことができない。
- ・ 操作方法によっては従来型の記事検索よりタイムロスが多いようだ。

そこで、演算子（AND, OR, NOT）を使用した検索を行いたい場合や、検索範囲（見出しのみ、など）を指定できるように、「ナビ型記事検索」では「検索オプション」を設けている（図10の囲み枠参照）。また2014年12月には利用者の声を参考にして、以下の対応を行った。

図10 検索オプションの設定

- ・ 期間選択のスライダーを両端から動かせるようにして操作性を改善
- ・ 画面での分類語・主題語の説明を強化
- ・ 絞り込みキーワード候補を4列表示にして横スクロールを廃止

5. 「ナビ型記事検索」から他メニューへの展開

「ナビ型記事検索」で実現した自動記事分類を使った新しい検索用インデックス生成処理は、他のメニュー・機能でも活用が始まっている。

5.1 直近の新展開：クリッピングへの応用

新しい検索用インデックス生成処理は、日経各紙だけでなくすべての媒体に適用されている。この結果、各媒体をまたがって、品質が一定した共通のタグを検索用インデックスとして付与できているので、同じ検索でも媒体間のブレを吸収することができ、一括検索の精度が格段に向上している。

この特性を生かして、クリッピングできる媒体の数を一気に増やしている（2014年12月時点で90媒体に対応済み）。

クリッピングとは、各媒体の最新の記事データが日経テレコンに収録されるタイミングで、利用者があらかじめ登録したキーワードにヒットする記事があればメールでお知らせを受け取ることができる機能である（クリップメール機能）。

5.2 今後の計画：業界サマリ

日経テレコンの業界情報メニューの中に、「業界サマリ」というサブメニューがある（図11）。このメニューでは、利用者が調べたい“業界”と“テーマ”（記事が扱う話題）を選択するだけで、その業界に関する情報を記事検索データベースから自動的にピックアップして提供している。「自社や取引先の業界の現状と今後の展望などの全体像を把握する」といった

図11 日経テレコン 業界サマリメニュー

ニーズに素早く応えるメニューである。

2015年にはこの業界サマリに「ニュースページ」を追加する予定である。最新の新聞記事やニュースから、業界や主要企業の日々の動きをウォッチできる機能で、次々と発生する最新情報を、業界別、トピックス別、企業別に自動的に仕分けるところに、自動記事分類技術を応用する予定である。

6. おわりに

「ナビ型記事検索」をスタートしたときに「迷わない」というキャッチコピーで新聞広告を出した。「知りたい情報へ、もう迷わない、迷わせない」というコンセプトであったが、「特別な検索スキルがなくても自然な流れで欲しい情報にたどり着ける検索メニュー」の探求はまだまだ始まったばかりだと認識している。

今後は、ナビゲーションの精度を高めていくことはもちろんであるが、利用者が欲しい情報にたどり着いた後の意図も理解して、関連情報を提示したり、新たな気づきを与える情報に導いたりするレコメンド機能も有用だと考えている。

また、タップ操作が中心となる携帯端末向けサービスでは、タグを生かしたインターフェースでモバイルワークスタイルならではの情報収集を実現するサービスを提供していきたい。

本文の注

注1) 文章を意味のある単語に区切り、辞書を利用して品詞や内容を判別すること。

注2) 日経テレコンヘルプサイト ナビ型記事検索・分類語一覧. <http://t21help.nikkei.co.jp/reference/cat845/post-563.html>

参考文献

- 1) 「日経テレコム」きょうスタート、生の経済情報を瞬時に. 日本経済新聞. 1984-04-02, 朝刊, p. 7.
- 2) 末吉行雄. NEEDS,日経テレコンの思い出(〈連載〉オンライン情報検索:先人の足跡をたどる(5)). 情報の科学と技術. 2008, vol. 58, no. 8, p. 408-414.
- 3) 出口信吾. 利用者の変化に合わせた新聞記事サービスへの取り組み. 専門図書館. 2011, no. 250, p. 16-22.
- 4) 神尾達夫. CTSを利用した新聞記事の画像データベース作成. 情報管理. 1989, vol. 31, no. 11, p. 933-941.
- 5) 神尾達夫. 新聞記事データベースにおけるキーワード自動抽出. 情報管理. 1989, vol. 32, no. 4, p. 283-293.

Author Abstract

Nikkei Telecom, providing business information via the Internet, released new search interface "navigation type article search" that introduced a new technology called automatic article classification process. Compared to full-text search, the new search significantly improves the relevance ratio of the search results, and makes the user to narrow down the search results while reflecting their intentions. In this paper, while introducing "navigation type article search", I describe the overview of automatic article classification techniques.

Key words

Nikkei Telecom, article, search, automatic classification, navigation type, language analysis, language understanding, keyword, thesaurus, index