ニュースディレクトリの構築と検索のための方式

Methods for Construction and Searching of News Directory Systems

劉 斌[†] ファム ヴァンハイ[†] ジェグ ケン[†] 徳田 雄洋[†]
Bin LIU Pham Van HAI Ken GEGOUT Takehiro TOKUDA
[†] 東京工業大学大学院情報理工学研究科
Graduate School of Information Science and Engineering, Tokyo Institute of Technology
{ryuu, hai, ken, tokuda}@tt.cs.titech.ac.jp

地球上の２０ヵ所以上のニュースサイトの４ヶ国語の自然言語で記述されたニュース記事の索引情報を毎日自動収集し，従来の Google ニュースなどのキーワード検索をベースとするニュース検索では検索が困難だった場合のニュース検索も一部可能とするニュースディレクトリシステムの構成法と検索法を提案する．従来のキーワード検索をベースとする検索では，母国語だがそのことの呼び名がわからない場合や知らない外国語で検索用語が全くわからない場合などの検索ができなかった．本システムでは，検索で数を絞り込んで，ユーザの知らない外国語のニュース記事の場合でも，別の自動翻訳無料サービスなどを利用し，記事の内容を読むことが可能となる．

1 はじめに

インターネットの普及につれて，ネット上で国内・国外のニュースをさまざまな自然言語で発信する新聞社，通信社，テレビ局などのニュースサイトが多数出現し，さらに Google ニュースなどのニュース検索サイトもいくつか出現している．これらのニュースサイトやニュース検索サイトのほとんどはキーワードをベースとするニュース記事の検索機能を提供している．

また，多くのニュースサイトではおおざっぱな部門分類の機能を提供している．しかしながら，実際にニュース記事の検索を行う時に現状の検索方法では調べたいニュース記事がうまく探せない場合がしばしば発生する．例えば，従来のキーワード検索をベースとするニュース検索では，母国語だがそのことの呼び名がわからない場合や、知らない外国語で検索用語が全くわからない場合など，検索することができない．

本論文は，地球上の２０ヵ所以上のニュースサイトの４ヶ国語の自然言語で記述されたニュース記事の索引情報を毎日自動収集し，従来の Google ニュースなどのキーワード検索をベースとするニュース検索では検索が困難だった場合のニュース検索も一部可能とするニュースディレクトリシステムの構成法と検索法を提案する．

第２章以降の論文の構成は次の通りである．第２章では提案するディレクトリシステムの構造および特徴について述べる．第３章ではニュース記事収集及びディレクトリへ記事を配置する方法について説明する．第４章ではディレクトリシステムを用いたニュース記事検索法を説明する．第５章では本論文の検索方法と従来のニュース記事検索法との比較を行う．第６章でまとめと今後の課題について論じる．

2 ニュースディレクトリシステム

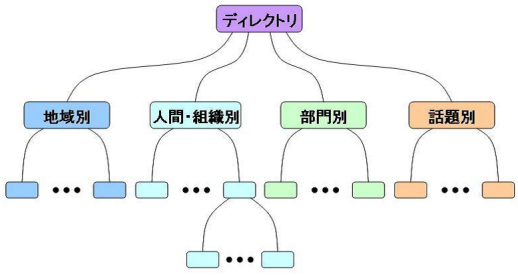

図 1: ディレクトリ構造

2.1 ディレクトリの構造

提案するニュースディレクトリシステムは英語，日本語，フランス語，中国語４種類の言語に対応し，ニュース記事の内容からニュース索引情報を分類し，保管する大きなリポジトリである．ディレクトリ全体はツリー構造となっており，一番上は記事を類別する４つの分類法がある．それぞれは地域別，人間・組織別，部門別と話題別である．これらの分類法の下でさらにそれぞれの基準でニュースを細分化する．

1. 地域別

   地域による分類はまず一般的な分け方で地球上を大きくいくつかのエリアに分け，その下に該当エリアに含まれる国・地域などを地域別にサブグループを作り，さらにその下に日付でグループ分けし，中にニュース記事へのリンクを配置する．この分類法はニュースが言及した内容から地理的空間上で記事を分類する方法である．

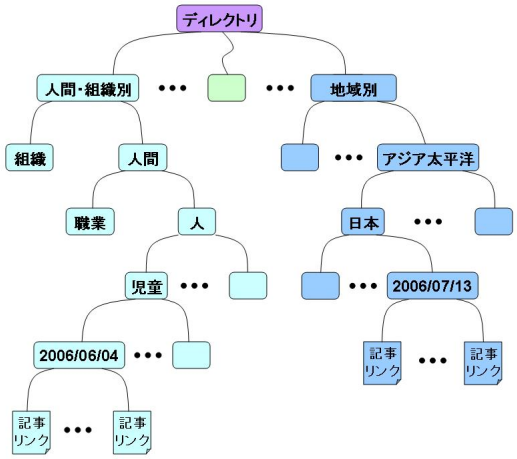

図 2: 地域別と人間・組織別

2. 人間・組織別

   人間・組織による分類はまず大きく「人間」と「組織」２つのグループに分ける．「人間」の下に「人」と「職業」２つの分類基準を設け，それぞれの基準でさらに人間を細分化し，日付で記事へのリンクをグループ分けする．「組織」の下に「企業」と「国際機関」２つのサブグループを設け，それぞれの基準でさらに組織を細分化し，日付で記事へのリンクをグループ分けする．この分類法はニュースの言及内容から記事を人間社会上の分け方で分類する方法である．

3. 部門別

   部門による分類はまず大きく「一般」,「ビジネス」,「技術・科学」,「健康」と「スポーツ」の５つのグループに分ける．それぞれのグループはさらに細分化できるレベルまでサブグループに再分類し，最後に日付で記事へのリンクをグループ分けする．この分類法はニュースの内容から判断し，言及内容の分野で分類する方法である．

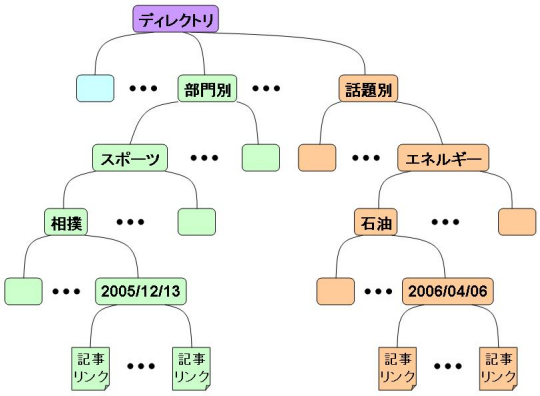

図 3: 部門別と話題別

4. 話題別

   話題による分類はまずいくつか大きな話題の分野を用意し，その下にさらに細かく各話題のサブグループを作り，中には日付で記事をグループ分けする．この分類法はよく話題に出るテーマでニュース記事へのリンクを分類する．

2.2 ディレクトリシステムの特徴

提案するディレクトリシステムは以下のような特徴を持つ．

2.2.1 動的なディレクトリ

ニュースは刻々変化するダイナミックのものであるため，時期により盛り上がる話題のニュースもあるのに対し，時間の経過につれて触れなくなるテーマもある．分類に登録されるニュース記事リンクの登録頻度により，ディレクトリの構造が動的に変わり，ニュースの言及変化を反映することができる．

「その他」分類

ディレクトリのいくつかの階層に「その他」という特別な分類が設けられている.「その他」分類は分類不明、またはその分類に入る分類の記事数が少ない，または出現回数や頻度が低い記事の分類先である．動的変更のための基本頻度データとしてどのような頻度データを使うかは現時点では検討中である．その１つの方法として，各分類にニュース記事リンクを登録するとき，その時点から１日前までにこの分類に登録された記事リンク数，１ヶ月前までにこの分類に登録された記事リンク数，１年前までにこの分類に登録された記事リンク数とその分類に登録された記事リンク総数４つの値が得られる．分類に登録される記事リンクの登録頻度と記事リンクの蓄積数を考慮し，４つのパラメータに異なる係数で重み付けし，足し合わせ得られた値で該当分類を「その他」分類に統合（降格）するかまたは「その他」分類から展開（昇格）するかを判断する．

このように，ニュース記事の出現頻度と記事総数に応じ，システムが世間の話題の変動に追付いているように動的に自身の構造を変える．

2.2.2 多言語への対応

システムは４種類の言語に対応している．ディレクトリに存在する各分類（ノード）は，記事の使用言語によらずにニュース記事索引情報を集める．これを実現するために各分類にキーワードリストを付属させる．

キーワードリスト

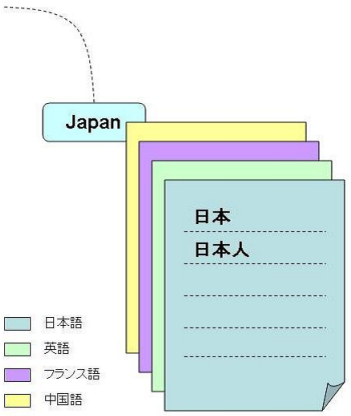

図 4: キーワードリスト

ディレクトリにある各分類をノードとし，各ノードに対応するキーワードリストをシステムの構築時に用意する．該当分類を特徴づけられるすべてのワードをキーワードリストに登録する．１つのリストに複数のエントリーがあり，エントリー間は OR 関係で記事を特徴付ける．登録する際に，AND 条件で該当分類を特徴付けるワードを１つのエントリーに登録し，OR 条件で該当分類を特徴付けるワードをそれぞれ１つのエントリーに登録する．このように，各分類に記事を配置する基準を作成する．

各ノードに共通のキーワードリストを言語毎に作成し，ノードに付加する．このように，異なる言語で記述されたニュース記事の索引情報が集めてきても，それぞれの言語に対応するキーワードリストを使えば，索引情報の内容判定が行え，類別できる．

3 記事の収集と配置

提案したニュースディレクトリシステムは毎日ニュースサイトから当日のニュース記事の索引情報をローカルに集め，記事の標題と内容で判断し，ディレクトリに対応した分類の中に配置する。

3.1 記事索引情報の収集

システムは現在，下記のような世界中から２４個のニュースサイトから毎日その日のニュース記事の索引情報を集めている．言語は４種類ある．

地域名	サイト URL	使用言語
オーストラリア	http://www.smh.com.au	英語
バングラデシュ	http://www.thedailystar.net	英語
中国	http://www.xinhuanet.com/english/index.htm	英語
	http://www.sina.com.cn/	中国語
ドイツ	http://www.dw-world.de/dw/	英語
フランス	http://www.lemonde.fr/	フランス語
イラン	http://www.tehrantimes.com/	英語
イタリア	http://www.agi.it/english/news.pl	英語
	http://www.lifeinitaly.com/news/Default.asp	英語
日本	http://mdn.mainichi-msn.co.jp	英語
	http://www.yomiuri.co.jp/dy/index.htm	英語
	http://www.asahi.com/	日本語
クウェート	http://www.kuwaittimes.net/	英語
フィリピン	http://newsinfo.inq7.net/index.php	英語
	http://www.manilatimes.net/	英語
ロシア	http://english.pravda.ru	英語
シンガポール	http://www.channelnewsasia.com	英語
韓国	http://english.yna.co.kr	英語
台湾	http://www.chinapost.com.tw	英語
タイ	http://www.bangkokpost.com	英語
トルコ	http://www.zaman.com	英語
アメリカ	http://www.nytimes.com	英語
ベトナム	http://www.vnanet.vn/Home/tabid/117/Default.aspx	英語
	http://www.thanhniennews.com/	英語

図 5: サイトリスト

ニュースの収集部分を担当するクローラは各ニュースサイトのトップページから開始し，ページソース

からリンク情報を抽出し，さらにそのリンクから展開させる．この操作の反復でニュースサイトのドメイン上すべてのページ URL を獲得することができ，さらに URL のパターンなどで当日の記事かどうかを判断し，記事ページの索引情報を収集する．索引情報をローカルに保存し，これらを区別する URI として元の記事ページの URL を利用する．

3.2 ディレクトリへの配置

記事ページへのリンクをローカルに保存したあと，ディレクトリへに配置する．記事ページのソースから記事内容のテキスト情報を抽出し，各言語の形態素分析ツールを用いて記事内容のテキストを形態素単位に分割する．このように，記事毎に形態素の集合が得られる．さらに，分類のキーワードリストを用いて形態素集合の要素調べ，ディレクトリのどの分類に含まれるかを判定する．最後にヒットした分類に該当記事の URI を登録する．

4 ディレクトリを用いた検索法

4.1 分類間の操作による検索

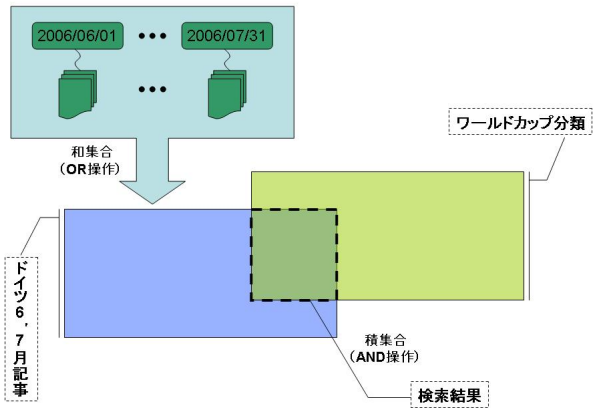

図 6: 分類間の操作による検索

提案したニュースディレクトリシステムは４つの分類方法を使い，ニュースの索引情報を複数の基準で細かく類別し，保管している．この構造上の特徴を利用し，分類と分類との和集合や積集合をとることで，簡単かつ正確に検索の結果を求めることができる．例として,「2006 年 6 月から 7 月までドイツにあったワールドカップのニュース記事」の検索を行うとき，まず地域別の分類法から 2006 年 6 月から 7 月までの間にドイツにあったニュース記事の和集合をとり，さらにこの和集合を話題別の分類法にある「ワールドカップ」分類に登録されているニュース記事の集合との積集合をとり，とれた積集合に含まれるニュース記事は検索条件に合った「2006 年 6 月から 7 月までドイツにあったワールドカップのニュース記事」となる．

4.2 言語間の検索

システムは複数の言語に対応しているため，１ヶ国語のみを使用し，他の言語で書かれたニュース記事を検索することができる．例えば,「2006 年中にあった狂牛病についての記事」を検索するとき，ユーザはディレクトリに対して日本語で検索条件を出し，これはディレクトリの狂牛病に関する分類に付加されている日本語のキーワードリストにヒットする．そこで，この分類に入っているすべての記事から 2006 年分の記事だけを抽出して返すと，英語の記事，日本語の記事，フランス語の記事と中国語の記事が全部ユーザの手元に届けられる．これは，日本語リストのほか，英語，フランス語，中国語のキーワードリストも普段それぞれの言語で記述された狂牛病に関する記事を集め，この分類の下に記事リンクを配置しているためである．

5 検索法の比較

提案したニュースディレクトリシステムでニュース記事検索を行うときに，複数の分類方法でニュースのリポジトリを構成したため，キーワード不明の時にでもディレクトリが提供している分類の階層を辿って，範囲を絞りつつ目的のニュース記事に極めて接近することができる．これは現在のニュースサイトが提供しているキーワード検索とニュース分類では行えない検索である．さらに，ディレクトリ構造上の特徴を利用した検索方法は検索結果の正確率を維持しながら，従来のキーワードをベースとする検索法より効率的な計算処理が期待できる．

6 まとめと今後の課題

本論文は，ニュース記事索引情報を複数の分類法により類別し，蓄積するリポジトリであるニュースディレクトリシステムとその特徴にあわせた検索方法を提案した．ディレクトリは最初の構築時に分類

や，キーワードリストを手動で作るだけで，システム動作開始後に複数のニュースサイトから毎日に新たな記事の索引情報を自動的に収集する．集めた索引情報をさらに分類し，動的に実社会の話題言及をディレクトリの構造上に反映させる．ディレクトリを用いた検索も，現在のニュースサイトのキーワード検索などの方法より多くの検索場面で使用でき，さらに，複数の言語に対応しているため，他国語で書かれたニュース記事の検索も可能である．

今後の課題としては，ディレクトリシステムの本格的実装，効果的な検索アルゴリズムの開発，動的ディレクトリの自動変更方式の改善，索引情報を収集するニュースサイトの追加などが考えられる．

参考文献

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang: ”Topic Detection and Tracking Pilot Study Final Report”, In Proceedings of the Broadcast News Transcription and Understanding Workshop, pp194-218, 1998

[2] BBC Sports: ”http://news.bbc.co.uk/sport”

[3] Courrier International: ”http://www.courrierinternational.com/gabarits/html/default_online.asp”

[4] Norberto Fernández-Garcá and Luis Sánchez-Fernández: ”Building an Ontology for NEWS Applications”, ISWC 2004

[5] F. Fornari. C. Monticelli, M. Pericolli and M. Tivegna, ”The Impact of News on the Exchange Rate of the Lira and Long-Term Interest Rates”, Economic Modeling, Elsevier, Vol. 19, No. 4 pp. 611-639, 2002.

[6] Google news: ”http://news.google.co.jp/”

[7] New York Times: ”http://topics.nytimes.com/top/reference/timestopics/index.html”

[8] K. Rajaraman and Ah-Hwee Tan:”Topic Detection,Tracking and Trend Analysis Using Self-organizing Neural Networks”, Pacific-Asia Conference on Knowledge Discovery and Data Mining , pp102-107, 2001

[9] Maria Vargas-Vera, Enrico Motta and John Domingue: ”AQUA: An Ontology-Driven Question Answering System”, The Open University 2004

[10] Yahoo news: ”http://headlines.yahoo.co.jp/hl”