3B3-3

メタデータを活用したNewsMLマネジメントシステムの試作

An Implementation of a NewsML Management System using Meta Data

児玉 政幸[*1] 大囿 忠親[*1] 新谷 虎松[*1]
Masayuki Kodama Tadachika Ozono Toramatsu Shintani

[*1]名古屋工業大学 大学院工学研究科情報工学専攻
Dept. of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology.

Recently, NewsML has been aiming at standardization for news management. A NewsML management system are strongly requested for effectively managing the NewsML. In this paper, we describe NewsML management system using metadata. Our NewsML management system is consist of three sub systems: the map search system, the related person search system, and the metadata editor. NewsML includes much and several metadata. We use metadata to manage geography data and personal data. By using these data, we can realize effectively the system and show the usability.

1. はじめに

近年，WWW 上での情報発信の即時性，及び情報共有の容易さにより，インターネットを介したニュース配信が盛んに行われている．それに伴い，ニュース管理／配信フォーマットの標準化が進められており，国際新聞電気通信評議会 (IPTC) が策定した XML ベースの NewsML[*1]を採用する配信社が増えている．そして現状，NewsML 開発環境の開発が積極的に取り組まれている [1]．NewsML の利用価値の高い特徴として，ニュースに様々なメタデータを付加できることが挙げられる．つまり，メタデータを有効的に活用することで，NewsML の知的な編集，管理，運用マネジメントサイクルを実現することが可能となる．しかし，現状，NewsML を対象とした開発環境は存在してはいるが，メタデータを有効的に活用していないという問題が挙げられる．したがって，メタデータを有効的に活用した NewsML システムを構築することが必要となってきている．我々はニュースの一連作業をマネジメントすることを目的とし，NewsML に含まれるメタデータを活用した NewsML 検索システム，及びメタデータエディタを試作した．本システムの対象ユーザはニュースを編集，投稿，管理する人達に限らず，ニュースを閲覧する側の人達もターゲットとしている．

以下，2. で NewsML マネジメントシステムの概要を述べ，3.，4.，5. で本システムの具体的な構成システムについて述べる．6. で関連研究を述べ，最後に 7. で本論文をまとめる．

2. NewsML マネジメントシステムの概要

我々が開発した NewsML マネジメントシステムは，NewsML のメタデータを活用することで，様々な検索環境を提供し，ニュース編集における検索作業を支援する．メタデータとして主に位置データ，及び人物データを用いる．これらのデータを用いることで，位置や人物を考慮したニュース検索が可能となる．具体的には本システムは，i) 地図を利用した検索，ii) 属性語と関連語を利用した関連人物検索，から構成される．

地図を利用した検索では，NewsML に含まれるメタデータである位置メタデータに着目し，地図を用いてニュースを検索したり検索結果を可視化することができる．また，補助機能としてニュース検索結果をグラフを用いて可視化することもできる．位置に着目して検索することで，いつどこで何のニュースが発生したかを直感的に把握することができる．

属性語と関連語を利用した関連人物検索では，NewsML に含まれる人物メタデータに着目し，関連人物に関するニュースを検索する．したがって，ユーザは目的のニュースを検索すると，システムが推薦した関連ニュースも手軽に閲覧することができる．

また，地図検索や関連人物検索では，位置データや人物データが NewsML のメタデータとして存在していることが前提である．しかし，メタデータ付加は人手で行う必要があり，入力を補助する枠組みが必要となる．本マネジメントシステムにメタデータエディタを組み込むことで，ニュース記事入力中にメタデータを自動的に抽出し，NewsML のメタデータ付加を支援することができる．したがって，メタデータエディタもニュース編集作業をマネジメントする必要不可欠なシステムであり，本マネジメントシステムの部分システムとする．

3. 地図を利用した NewsML 検索

3.1 概要

本システムは，JavaScript，PHP，及び MySQL を用いて構成される．本システムで利用する地図は GoogleMaps である．GoogleMaps は Google[*2]によって無償で API[*3]が公開されている地図検索サービスである．GoogleMaps を用いる理由は，i) 地図上に任意のオブジェクトを配置可能，ii) 地図の拡大縮小機能，及び地図タイプの切替機能などのインタフェース操作における充実性，の 2 点である．本システムでは NewsML に含まれるメタデータに基づき，地図を用いて可視化することで，視覚的な検索環境を実現する．図 1 に本システムのスナップショットを示す．

本システムでは地図を用いて直感的にニュースを検索することができる (以下，地図検索)．ユーザは地図上から興味のある地域をマウスで範囲選択することにより，選択範囲内のニュースを検索することができる．地図検索では，従来のテキストベースの検索エンジンとは異なり地理名を入力する必要がない．つまり，ユーザにとって地理名が未知な場所であっても，その場所で発生したニュースを簡単に検索できる．また，カテゴリセレクタでカテゴリを選択することで，同選択範囲内

連絡先: 児玉 政幸，名古屋工業大学大学院情報工学専攻，kodama@ics.nitech.ac.jp

*1 http://newsml.jp/

*2 http://www.google.com/

*3 http://www.google.com/apis/maps/

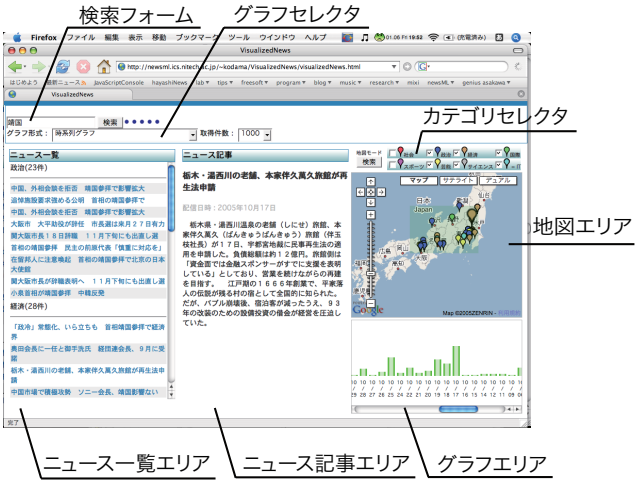

図 1: 地図を利用した NewsML 検索システムのスナップショット

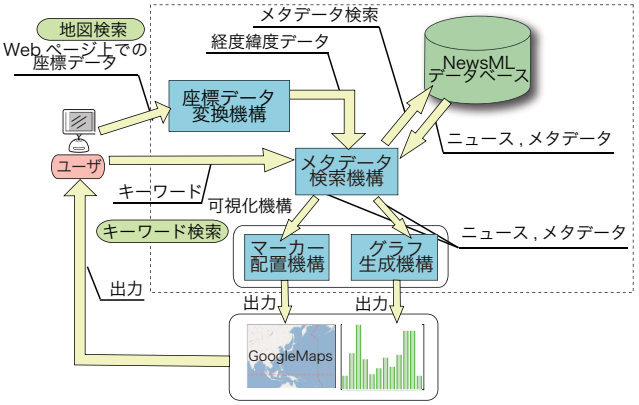

図 2: システム構成

で発生したニュースを再検索することができる．本システムで可視化されたニュース検索結果は，図 1 の地図エリア，グラフエリア，及びニュース一覧エリアに表示される．マーカー，及びニュース一覧エリアのニュースを選択すると，ニュース記事がニュースエリアに表示される．

本研究では，NewsML 検索結果を視覚的に表示する手法を 3 つ提案する．地図を用いた可視化では，地図検索の検索結果を可視化する．NewsML のメタデータである位置メタデータを活用し，地図上にマーカーを配置する．時系列グラフを用いた可視化では，キーワード検索の検索結果を可視化する．NewsML のメタデータである配信時刻情報を活用し，日付を横軸に，ニュース件数を縦軸にとった棒グラフを表示する．都道府県別グラフを用いた可視化では，キーワード検索の検索結果を可視化する．NewsML のメタデータである位置メタデータを活用し，各都道府県を横軸に，ニュース件数を縦軸にした棒グラフを表示する．

3.2 システム構成要素

本システムは座標データ変換機構，メタデータ検索機構，及び可視化機構から構成される (図 2)．さらに，可視化機構はマーカー配置機構，及びグラフ生成機構に分類できる．

座標データ変換機構は，地図検索の際呼び出される機構である．NewsML に含まれる位置メタデータは経度緯度で管理されているが，GoogleMaps の仕様上，選択範囲の経度緯度データを直接取得することはできない．したがって，本研究では Web ページ上の座標データから経度緯度データへの変換を行うことで，選択範囲の経度緯度データを取得する．本研究では，スタイルシートの zIndex を操作し，地図の上にレイヤーを貼ることで，Web ページ上の座標データ抽出を可能とした．抽出された座標データは地図エリアの各頂点の経度緯度データを用いることで，経度緯度データに変換することが可能となる．地図エリアの各頂点の経度緯度データは，GoogleMapsAPI を利用することで取得することができる．

メタデータ検索機構は，ユーザの検索要求から検索すべきメタデータを判断し，NewsML データベースからメタデータを検索する機構である．ユーザがニュースを検索すると，メタデータ検索機構は NewsML データベースに存在する適切なメタデータに対して検索を行う．メタデータ検索機構は，検索されたニュース，及びメタデータを本システムの可視化機構に渡す．

マーカー配置機構は検索された位置メタデータを基に，マーカーを地図上に配置する機構である．また，現在閲覧しているニュースが地図上のどのマーカーであるかを一目で把握するために，マーカーの拡大表示を実装した．ユーザがニュース一覧エリアのニュースタイトル，及び地図上に表示されるマーカーをクリックすると，マーカーが拡大表示される (図 1)．さらに，ニュースのカテゴリ情報を参照することで，マーカーをカテゴリ毎に色分け表示する (図 1)．上記処理は GoogleMapsAPI を利用することで可能となる．

グラフ生成機構は検索されたニュースを棒グラフを用いて可視化するための機構である．生成するグラフは，時系列グラフ，及び都道府県別グラフの 2 種類である．ユーザからの検索キーワード，及び NewsML のメタデータである位置メタデータ，及び配信時刻情報を組み合わせることでグラフを生成する．

4. 属性語と関連語を利用した関連人物検索

4.1 概要

本システムは，検索クエリとして入力された人物名から，関連する人物とその人物名を含むニュース記事を検索するシステムである．ニュースに含まれる属性語，及び関連語に着目し，関連人物に関するニュースを検索する．属性語は，格助詞の「の」で直接係る語,「=」で結ばれる語，及び「人物名 ()」の括弧の中身など，人物を直接形容する語を抽出する．関連語は，属性語以外で，同じ文中に共起する語を抽出する．抽出した属性語，及び関連語により，人物ごとに単語ベクトルを作成する．このとき，属性語と関連語に同じ語が含まれていても別の単語として区別する．そして，単語ベクトル間の類似度を計算することで，関連人物を検索する．

図 3 に関連人物検索システムの構成を示す．ユーザは，検索クエリとして人物名を入力する．システムは，NewsML から抽出した人物メタデータを活用することで，入力された人物のニュース，及び関連人物のニュースを返す．図 4 に実行時のスナップショットを示す．

4.2 関連人物検索手法

関連人物の検索手法について述べる．抽出機構により NewsML データベースのメタデータから人物名を得て，その人物名の出現するニュース記事から人物の属性語と関連語を抽出し，これをメタデータとして NewsML データベースに追加する．人物ごとに，属性語と関連語を人物データベースに格

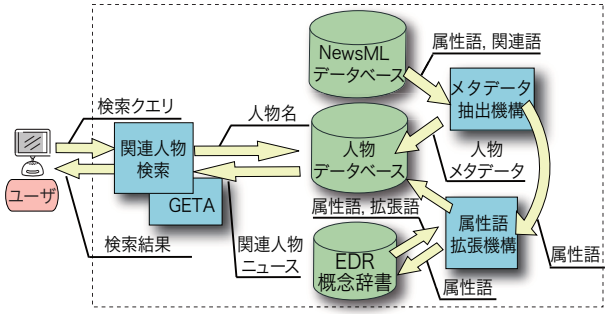

図 3: システム構成

図 4: 関連人物検索実行イメージ (検索語：小泉)

納する．その際に，属性語は属性語拡張機構により，EDR 概念辞書*4を利用して拡張される．属性語の上位語，及びその上位語の出現数を元の属性語のそれぞれ $\frac{1}{2}$，$\frac{1}{4}$ として人物データベースに追加する．抽出した属性語，及び関連語により，人物ごとに単語ベクトルを作成する．そして，単語ベクトル間の類似度を計算することにより，関連人物を検索する．属性語を拡張することで，属性語「総理」から属性語「大統領」を検索できるようにした．また，語の表記に揺れを抑えることも期待できる．

類似度は Singhal[2] の方法を利用した．Singhal の方法は，文書の長さによらず適切な類似度を求めることができる方法である．単語ごとに検索語の TF，及び検索対象の TFIDF の積を計算し足し合わせ，文書の長さにより正規化を行う．次に計算式を示す．以下の式で，q は検索クエリ，t は検索クエリ中の語，d は検索対象の文書，$wq(t|q)$ は語 t の q における重要度，$wd(t|d)$ は語 t の d における重要度，$norm(d)$ は d を正規化する尺度，$sim(d|q)$ は q と d の類似度を表す．

$$wq(t|q) = \frac{1+\log(TF(t|q))}{1+\log(\mathrm{ave}(TF(q)))} \cdot IDF(t) \quad (1)$$

$$wd(t|d) = 1+\log(TF(t|d)) \quad (2)$$

$$norm(d) = (\mathrm{ave}(\mathrm{len}(d)) + 0.2 * (DF(.|d) - \mathrm{ave}(\mathrm{len}(d)))) \cdot \left(1+\log\frac{TF(.|d)}{DF(.|d)}\right)$$

$$sim(q|d) = \frac{1}{norm(d)} \cdot \sum_{t\in q}(wq(t|q)\cdot wd(t|d)) \quad (3)$$

(4)

*4 http://www2.nict.go.jp/kk/e416/EDR/

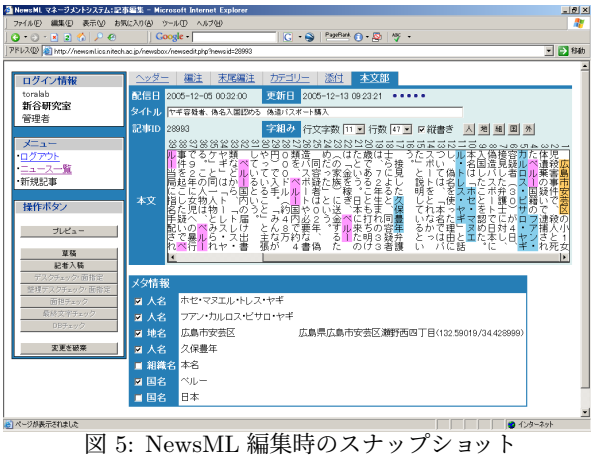

図 5: NewsML 編集時のスナップショット

5. メタデータエディタ

5.1 概要

メタデータエディタでは，ニュース入力中にメタデータを自動的に抽出するシステムである．抽出する具体的なメタデータは，地名，国名，人名，組織名，及び外国の地名である．メタデータが抽出されると，その記事に付加するメタデータ候補が表示される．ユーザが候補からメタデータを選択することで，システムは NewsML のメタデータとして自動付加する．また，システムが推薦したメタデータが間違っている場合，マウスで記事の抽出対象部分を選択し，正しいメタデータを指示することで，解析が行われ正しいメタデータが付加される．図 5 に NewsML 編集時のスナップショットを示す．

5.2 メタデータの自動抽出手法

図 6 にメタデータ抽出の流れを示す．まず，ニュース記事に対して CaboCha により形態素解析，及び固有表現抽出を行う．次に形態素解析結果に対して，パターンマッチングを行い，メタデータ付加の対象となる単語列を抽出する．例えば，固有名詞-地域が連続していれば，住所として単語列を抽出する．抽出した単語列の種類に応じて，データベースを検索し，メタデータ付加のための情報を生成する．住所の場合，住所データベース (住所と座標の対応表，街区レベル位置参照情報*5を利用) を検索し，住所から座標データを得る．住所と座標データの組を位置メタデータとして，Web ブラウザに返す．人物名の場合，人物名を一意に特定する URN を人物データベースから検索し，人物名と URN の組をメタデータとして Web ブラウザに返す．メタデータの確認が行いやすいように，住所には地図へのリンク，人名には同一人物が参照されている記事へのリンクが付けられる．また，メタデータ管理機構により修正結果が蓄えられ，辞書を強化することで抽出精度が向上する．

6. 関連研究

地理データは Web 上の情報の組織化において有用な情報とされている．村山 [5] は住所情報を Web 上から抽出して，メ

*5 http://nlftp.mlit.go.jp/isj/

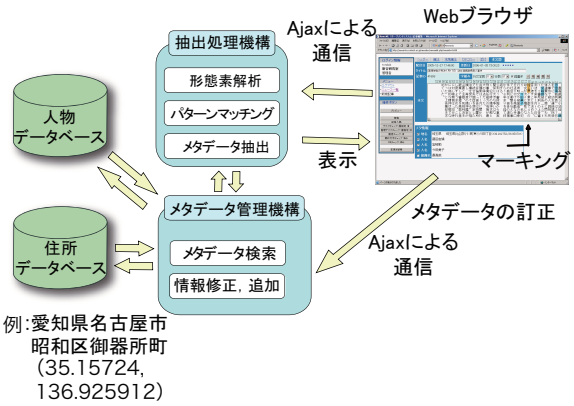

図 6: メタデータ抽出の流れ

タデータとして利用するデータベースを生成する手法について述べている．また，地理データとニュースを融合させた可視化システムに BuzzTracker[*6]がある．BuzzTracker は，ニュースの発生地域間における関係性，及びニュースに含まれる地理データを可視化している．他の機能として，地図上にニュースを配置する際，ニュース件数に比例した大きさの円を用いて，ニュースの注目度を視覚的に表現している．Crime map[*7]は，シカゴで発生した過去の犯罪場所を地図上に可視化するシステムである．本研究の地図を用いた NewsML 検索システムは，NewsML に含まれるメタデータを活用することで，単に地図上にニュースを可視化するだけでなく，グラフを利用した統計情報もインタラクティブに取得することができる．

人物データを抽出する研究として西野 [6]，山本 [7]，及び森 [8] の研究がある．西野 [6] は職種や業種を手がかりにして人物名や企業名を，あるいは人物名や企業名を手がかりにして職種や業種の情報を獲得する手法について述べている．山本 [7] は Web 上から知りたい人物の職業の人名リストを収集し，その人名リストに対して表解析を適用することで，人物データを抽出している．本研究で実装した関連人物検索は対象となる人物の属性情報を用いることで，その人物に関連した人物のニュースを検索することができる．ここでの属性情報とは，ニュース中に出現する人物に付随した情報を指し，職種，年齢，所属などを含む．本システムでは，抽出だけにとどまらず NewsML のメタデータとして保持することで，NewsML の編集や検索に再利用することができる．

メタデータはインターネットの普及により，WWW を中心とするインターネット上のリソースを記述する枠組みとして注目を集めている [3]．本研究が対象としている NewsML には非常に多くのメタデータを記述することが可能である．例えば，記事の種類，作成者，更新時刻，重要度，関連記事，提供者，記事の内容，及び場所などがそうである．その反面，上記メタデータの付加は人手で行う必要があり，ニュース編集者にとって大きな負担となる．メタデータを有効的に扱うには，メタデータ付加を補助する仕組みや，メタデータを自動的に付加する仕組みが必要不可欠である．メタデータを自動付加しようとする研究として本間 [4] の研究がある．本研究で開発したメタデータエディタは半自動的にメタデータを付加することで上記問題を解決している．半自動的に行うことにより，間違っているメタデータを簡単に修正することも可能である．

*6 http://www.buzztracker.org/
*7 http://www.chicagocrime.org/map/

7. おわりに

本論文では，NewsML のメタデータを活用した NewsML 検索システム，及びメタデータエディタを実装した．地図を用いた NewsML 検索システムでは，位置メタデータと地図を組み合わせることで，ニュースを視覚的に検索，及び閲覧することができる．関連人物検索では，人物メタデータを用いることである人物に関連した人物のニュースを検索することができる．そしてメタデータエディタでは，メタデータを半自動的に処理することで，ユーザのニュース編集作業における負担を軽減することができた．また，NewsML フォーマットに自動変換することで，NewsML の内部構造をユーザが把握する必要がなくなり，NewsML の予備知識を持っていない人でも編集可能である．

以下に今後の課題を挙げる．現在のメタデータエディタでは，ランドマークを抽出することができない．ランドマークはニュース記事に頻繁に出現する語であり，ニュースの発生場所を特定するための重要なファクターである．さらに，ランドマークを抽出できれば，地図検索でもその情報を基にニュースを検索することが可能となり，より必要性があると考えられる．関連人物検索では，辞書強化による人物抽出精度の向上，類似度計算方法の改良，及び関連語抽出の精度向上があげられる．関連人物検索において，不適合である検索結果のほとんどが人物抽出誤りによるもので，抽出精度をあげることは，そのまま適合率をあげる事につながる．さらに，本システムの信頼性を評価するために，実際にニュース編集に携わる人々に本システムを使ってもらい，システムの有効性を調査する必要がある．

参考文献

[1] 井上明，猪狩淳一，金田重郎，“ニュース配信のための国際データフォーマット NewsML:その概要と現状について”，情報処理学会論文誌，Vol.2002，No.056，2002

[2] Singhal A.，Buckley C.，and Mitra M.，“Pivoted document length normalization”，In Proceedings of SIGIR’96，pp.113-126，1997

[3] 浦本直彦，津田宏，上田隆也，佐藤研治，野村浩郷，“WWW に見るメタデータの標準化動向”，情報処理学会研究報告「データベース」，Vol.1998，No.34，1998

[4] 本間克哉，“メタデータ作成支援プログラムの開発”，日本測量調査技術協会，第 25 回 技術発表会論文特集，APA No.85-10，2003

[5] 村山紀文，南野朋之，奥村学，“メタデータ付与のための住所録自動生成”，情報処理学会研究報告「自然言語処理」，Vol.2004，No.73，2004

[6] 西野文人，落谷亮，“新聞記事からの人物・企業情報の抽出”，情報処理学会研究報告「自然言語処理」，Vol.1998，No.81，1998

[7] 山本あゆみ，佐藤理史，“ワールドワイドウェブからの人物情報の自動収集”，情報処理学会研究報告「知能と複雑系」，Vol.2000，No.3，1999

[8] 森純一郎，松尾豊，石塚満，“語の共起情報に基づく Web 上からの個人メタデータ抽出”，第 7 回セマンティックウェブとオントロジー研究会資料，SIG-SWO-A403-01，2004