2D4-2

ニュース記事に対する価値付けのための ワークフローマネジメントシステムについて

On a Workflow Management System for a News Agency to Valuate Articles

児玉政幸[*1] 伊藤正都[*1] 大囿忠親[*1] 新谷虎松[*1] 青崎保好[*2]
Masayuki Kodama Masato Ito Tadachika Ozono Toramatsu Shintani Yasuyoshi Aosaki

[*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, Nagoya Institute of Technology

[*2](社) 共同通信社
Kyodo News

A news agency distributes a news article with its metadata to subscribers. Subscribers can easily choose valuable news by using metadata, which include information of news value, category, and so on. Such metadata can be used to reuse archived news articles effectively. Provided information of news value by a news agency is a rank of a day. Much writers are requested to attach to articles. However making rich metadata is very costly. We propose a workflow management system for a news agency to support to conduct useful metadata. Our system is a multiagent system to coordinate editors, writer, subscribers and readers in order to reduce the cost of making metadata. KDML is a markup language for metadata of a news article. Agents integrate news value information collected by subscribers and readers with metadata and send them to writers as feedback. Our system can help writers to make effective metadata easily.

1. はじめに

通信社では，ニュース記事 (以降，記事と記す) を編集するために過去の記事を参照する場合や，過去のある期間における重要記事をリスト化する場合に，過去の記事を検索し再利用することがしばしばある．本稿では，ニュースは事象を表し，ニュースを文章などで表現したものを記事と呼ぶ．

現状の記事検索システムでは，記事重要度，記事中に出てくるキーワード，およびカテゴリなどのメタ情報によって記事を検索できる．記事重要度とは，新聞社に記事を配信する際に通信社が各記事に付けるメタ情報であり，記事選択のためのコストを抑えつつ，重要なニュースを逃さないことを目的とする．これは記事とともにアーカイブされ，効率的に重要なニュースを検索する際に用いられている．

しかし，記事重要度はその日に配信される短期的な記事群からの絞り込みには適しているが，過去数日にも渡る長期的な記事群からの絞り込みには適していない．例えば，一日に配信される記事のうち重要度が高く設定される記事はごくわずかであり，比較的低コストで重要なニュースを探すことができる．しかし，過去の記事中から重要ニュースを検索する場合，重要度が高く設定されている記事が多数あり，通信社が定めた３段階の指標ではさらに少数に絞り込むことは難しい．

絞り込みの作業は人手 (記者やデスク) によってなされているが，彼らは日々記事を作成する業務に追われており，関連する記事の検索や重要な記事のリストを作成するといった作業にコストをかけることは思わしくない．

本稿では，ニュースの重要度を決める価値付けについて議論し，記事への価値付けを容易にするにあたり，NewsML 文書間でのリンクの維持手法として NVML(News Value Markup Language)[1] を基にした，大囿らの KDML(Knowledge Discription Markup Language)[2][3] を用い，現状のニュース配信フローを応用して記事に様々な価値情報を付加するシステムを提案する．

連絡先: 名古屋工業大学大学院工学研究科情報工学専攻，〒466–8555 名古屋市昭和区御器所町，{kodama, itomasa, ozono, tora}@ics.nitech.ac.jp

1.1 ニュースの価値

ニュースの価値を絶対的に評価するのは困難であり，相対的な評価にならざるを得ない．現在は通信社によって１日の内での相対的な評価が行われており，その評価結果はメタ情報としてアーカイブされる．ニュースの相場観に基づき，プロの記者が判断した価値を配信，アーカイブすることは有効である．しかし，ニュースの価値は個人の主観に強く依存するため，あらゆる視点から相対的に評価を行い，様々な価値を記述できる形式にすることが望ましい．

本研究では，通信社の価値付け以外にも新聞社や読者からの価値を記述できるようにする．新聞社による価値付けとして記事の掲載状況を，読者による価値付けとして記事の閲覧状況を利用する．

2. メタ記事配信言語 KDML

ニュースに価値付けを行う際に必要な仕組み，KDML の導入に付いて述べる．

近年，ニュース配信・管理フォーマットとして，国際新聞電気通信評議会 (IPTC)[*1]により策定された，NewsML(News Markup Language)[*2]が利用されている [4]．NewsML とは XML ベースのフォーマットであり，柔軟に多くのデータを格納することが可能である．ニュース配信・管理における NewsML 文書の内部は，大きく次の二つに分けられる．一つ目は，記事自体の項目 (コンテンツ) であり，二つ目は記事間のリンク，構造を表す項目である．

一方，KDML とはニュース配信における NewsML 文書から，リンク管理情報とコンテンツ管理情報を切り離したものであり，NewsML 文書での配信・管理を独立して制御するための言語である．そのため，記事インンデックス (お勧め記事一覧) は KDML で表された記事の関係を基に作成を行う．

2.1 KDML を用いた記事管理

記事は頻繁に作成され，その都度配信される．また，通信社の社会的な制約として，一度配信した記事を訂正する記事の配

[*1] http://www.iptc.org/
[*2] http://newsml.org/

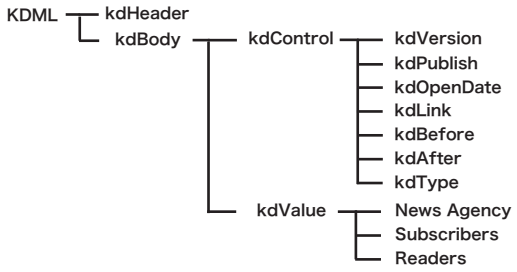

図 1: KDML におけるメタ情報構造例

表 1: kdControl の仕様

タグ	用途
kdControl	KDML 文書管理情報
kdVersion	KDML 文書バージョン
kdPublish	KDML 文書配信日付
kdOpenDate	記事解禁日時
kdLink	KDML 同士のリンク情報，KDML-NewsML 間のリンク情報
kdBefore	修正前情報
kdAfter	修正後情報
kdType	KDML の種類

信は可能であるが，一度配信した記事の実データは訂正する事ができない．しかしながら，NewsML 文書を用いたリンク管理では，NewsML 文書間のリンク情報がそれぞれの NewsML 文書に含まれるため，リンク整合性を管理する必要があり，その管理コストが大きくなる．また，記事の更新時には，更新前のリンク情報，および関連記事へのリンク情報を更新する必要があり，結果リンク情報は複雑になり，記事に対する価値付けは容易ではない．そのため，本章では価値付けのための記事インデックス作成・管理において，NewsML から配信・管理に特化した KDML を用いることで記事インデックスを管理する手法を提案する．

2.2 KDML の仕様

KDML におけるメタ情報の階層構造は図 1 である．KDML はヘッダと内容部があり，ヘッダ (kdHeader) にはその KDML 文書自身のメタ情報が包含される．内容部 (KdBody) には KDML 文書間のメタ情報 (kdContorol)，および KDML 文書で管理している NewsML 内の記事に関するメタ情報 (kdValue) が包含される．本稿において記事インデックスを作成するために用いる情報は KDML 文書間のメタ情報 (kdControl) であり，表 1 に仕様を示す．

kdControl におけるバージョン (kdVersion) とは KDML 文書の配信数である．常に最新のバージョンを持つ KDML をリンク構造の整合性管理するために利用する．KDML-NewsML 間のリンク管理情報 (kdLink)，は KDML 文書同士，もしくは KDML 文書と NewsML 文書を結びつける情報である．kdLink はリンクの構造を示しており，リンクの整合性を保つために最も重要な情報である．kdControl では，その他に配信日時 (kdPublish)，解禁日時 (kdOpenDate)，KDML 間，新規・更新の種別 (kdType)，修正前情報 (kdBefore)，修正後情報 (kdAfter) などの情報が含まれる．これらは KDML を用いてニュース配信・管理，記事インデックスの作成・管理を行うためのメタ情報群である．

表 2: リンク整合性管理の比較

	メリット	デメリット
NewsML	・NewsML のみで配信が可能	・記事と管理情報が混在し煩雑 ・検索コストが大きい
KDML	・文書間のリンク管理が容易 ・リンク構造の表示が容易 ・記事への価値付け	・KDML が必要 (データ量の増加)

2.3 記事リンク整合性機構

KDML を用いて記事リンクの整合性を管理する利点は，NewsML 文書リンク構造分離し，扱い易くすることである．NewsML 文書のみでのニュース配信では，NewsML を利用した記事の管理と記事が混在しており，リンク構造の取得が容易ではない．結果として，記事間のリンク構造から作成される記事インデックスの作成が容易ではなくなり，記事への価値付けに困難を生じさせる．記事リンク整合性機構を利用し，リンク情報を NewsML 文書からを KDML に分離する事で，記事ツリーの作成を従来の NewsML 文書から作成するのと比較し，容易にすることが可能となる．この結果，ニュース配信時にリンク構造を用いた，記事のリスト表示，記事への価値付け，記事インデックスの作成が容易になる．

2.4 リンク整合性管理の比較

表 2 において，KDML を用いたニュース配信と従来通りの NewsML でのリンク整合性管理方法のメリット・デメリットを比較した．

KDML によりリンクの整合性を維持する事は，配信コストを考える場合，NewsML 文書のみでのリンク整合性管理と比較し不利である．しかしながら，昨今のネットワークインフラ，ストレージの発展を考えるに，大きな問題ではないと考えられる．逆に，KDML を用いてリンク整合性を維持する事により，より容易に記事間の相関をリンクとして表現する事が可能となる．その結果，記事インデックスの作成が容易になり，価値情報に基づいた記事インデックスを作成するコストを減らす事が出来る．

3. 価値付けのためのワークフロー管理

通信社における記事配信において，掲載状況，および閲覧状況は現在管理されておらず，まずそれらのデータを収集する必要がある．現在の記事配信の仕組みを大幅に変えることは事実上不可能であるため，既存の記事配信フローを利用することが前提となる．

掲載状況，および閲覧状況を収集するために，新たにフローを追加する (図 2)．提案する追加フローは，記事インデックスファイルを共有することで実現される．記事インデックスとは，掲載・放送の確率が高いと予想される主要記事をまとめた出稿メニューである．記事インデックスは，KDML によって記述されており，各記事毎にそれに対するリンク，および各記事に関するメタ情報が記述されている．

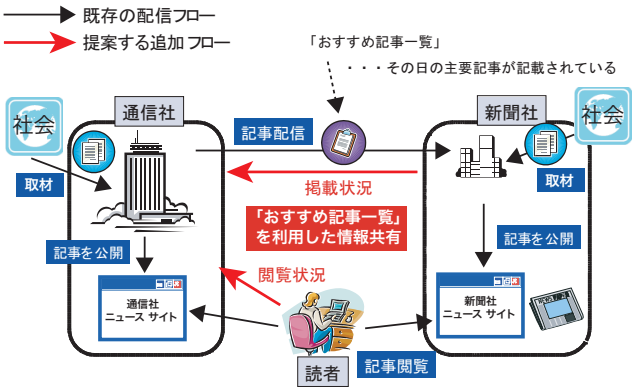

図 2: **提案する記事配信フロー**

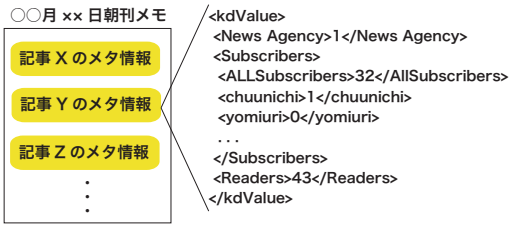

図 3: **価値情報の記述**

我々は記事インデックスに掲載状況，および閲覧状況のメタ情報を記述することを提案する．ここで，通信社による記事重要度，新聞社の掲載状況，および読者の閲覧状況を価値情報と称する．価値情報を共有するために，本研究では通信社側の記事配信サーバを中継サーバとして記事配信システムを構築する．掲載状況の取得に関して，従来は通信社が新聞社に対して記事をプッシュ配信していたが，互いのシステムが独立しているため配信した記事の掲載状況を通信社側で把握することは困難であった．本システムの従来との差分は，新聞社側が記事配信サーバのページから受信する点である．通信社は従来通り記事をプッシュ配信するが，配信先は新聞社ではなく記事配信サーバである．新聞社側にとっては配信記事を選択する場が Web ブラウザ上に変わるのみである．Web ページから記事を選択させることで，Web ブラウザから新聞社の選択状況などを把握できる．

3.1 価値情報の記述

価値情報 ⟨kdValue⟩ には通信社の記事重要度，新聞社の掲載状況，および読者の閲覧状況が含まれる (図 3)．通信社の記事重要度 ⟨NewsAgency⟩ では従来通りのタグ付けがなされ，数値の 0 は一般記事を，1 は最重要であることを，2 は重要であることを表す．新聞社の掲載状況 ⟨Subscribers⟩ には新聞社毎の記事選択の有無がタグ付けされ，数値の 0 は記事が選択されない場合に，1 は記事が選択された場合にタグ付けされる．読者の閲覧状況 ⟨Readers⟩ には読者の記事に対するアクセス数がタグ付けされ，その数値は通信社の Web サイト上にあるその記事をどのくらい多くの読者が閲覧しているかを意味する．

3.2 Web を利用した価値情報のタグ付け

新聞社が通信社からの配信記事に対して評価しタグ付けする作業は高コストである．また，読者が Web サイトの記事を閲覧し評価することは面倒であるため，データを収集すること

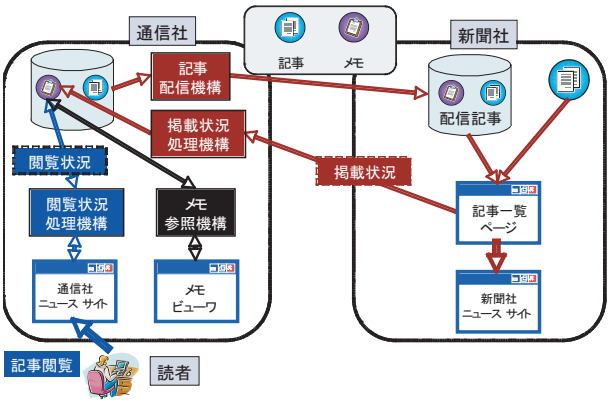

図 4: **価値付けのためのワークフロー**

は難しい．価値情報をメタ記事にタグ付けする際，極力負荷を与えないようにする必要がある．新聞社や読者に負荷を与えることなく価値情報をタグ付けさせることは重要であり，実運用を考えた上でも効率的である．

我々は記事の取得・閲覧状況を検知し，自動的にメタ記事のメタ情報を更新するシステムを試作した．図 4 にシステム構成を示す．本システムのインタフェースは Web ブラウザであり，システムプログラムは通信社側のサーバに全て用意する．

編集者が KDML を作成し配信可能状態にしておくと，システムが記事インデックスビューワに KDML の情報を反映させる．ここでシステム利用者の負荷を減らすために，記事インデックスビューワのページを常時立ち上げておくことで，ページ更新作業がなくても最新の情報をリアルタイムに閲覧できるようにした．これは Ajax 技術によるものである．新聞社は従来通りメタ記事の情報から配信記事を取捨選択し，自社の記事として利用する．システムは新聞社がどの記事を取得したかを常に監視し，メタ記事の ⟨Subscribers⟩ 要素中の値を更新する．これは記事の利用率を更新することと同義である．通信社のニュース Web サイトにもシステムを適用することで，読者の閲覧動向を監視させる．読者が記事を選択すると，システムはその記事のメタ情報を即座に更新する．これは記事の閲覧率を更新することと同義である．

4. システム構成

4.1 ドキュメントサーバ

ドキュメントサーバは，配信記事を保存，検索するシステムであり，1 台の検索マスタサーバと複数の検索スレーブサーバから構成される．検索マスタサーバの仕様は CPU：Pentium4 3.6Ghz，RAM：2GB である．また，検索スレーブサーバの仕様は CPU：Athlon64x2 2GHz，RAM：2GB である．

本システムを構築するにあたり，P2P 連携機能を持つオープンソース全文検索エンジン Hyper Estraier[*3] を利用した．P2P 機能によりインデックスを分散保持することが可能な Hyper Estraier を用いる事で記事検索に対するスケーラビリティーを高める事が出来る．

具体的に，検索マスタサーバのみでの検索の場合，600 万ドキュメントから検索を行う場合には検索ワードにもよるが 5 秒から 10 秒の時間が必要である．一方，分散検索システムに置き換えることで，検索時間を十分の一にする事が可能である．

*3 http://hyperestraier.sourceforge.net/

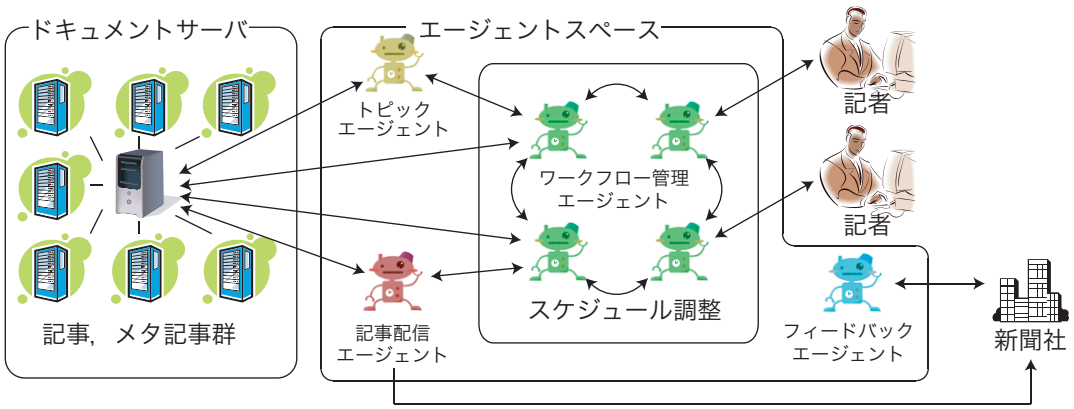

図 5: ニュースワークフロー

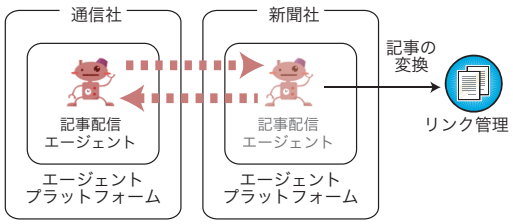

図 6: 記事配信エージェント

4.2 エージェントスペース

ワークフロー管理エージェントは，記者のマネジメント，ワークフロー管理を行うエージェントである．記者のマネジメント，ワークフロー管理とは，デスクからの取材依頼，記事に対するタグ付けなど記者への仕事を割り振る事である．あるエージェントは他のエージェントからの依頼を受け，業務に余裕のある記者に対して仕事を割り振る．余裕のある記者とは，現在執筆中でない記者などが挙げられる．タスクの割り振りは，各出稿部 (ex. 社会部，政治部，経済部) 内に限定する．出稿部内で均等にタスクを配分することで，部署全体の作業効率を上げることを目指す．

トピックエージェントは，ニューストピックの追跡を行うエージェントである．トピックエージェントは，記事へのタグ付け等を参考にし，記事間の関係を作成する．その際に，ドキュメントサーバに追加された記事，記事インデックスに関しての整合性，タグ付けの有無などを確認する．不整合，またはタグ付けが不十分な場合には，ワークフローエージェントにその事を伝え，記者に対する作業を要求する．

また，トピックエージェントは記事の分析に今後利用できる．例えば，重要視していなかった記事が読者に頻繁に閲覧されていることの発見，新聞社毎に掲載状況を計るほか，カテゴリや日付などを価値情報と組み合わせた統計情報の分析も可能である．本研究で取得する共有情報をトピック分析に応用することもできる．我々は記事群を各トピックに分類し，メタ情報として記事に付加するシステムを開発している [5]．掲載状況，閲覧状況をメタ情報として管理し，トピック内の各記事に相対的な重みを付けることで，トピック内での盛り上がりを表現することも可能である．

記事配信エージェントは，記事を配信するためのエージェントであり，拡張性の高い記事配信システムを構築するにあたり，モバイルエージェントが有効である．記事配信エージェントは配信記事と共に新聞社のエージェントプラットフォームへ移動する (図 6)．記事配信エージェントを新聞社側に送り込むことで，従来整合性のとれていなかった記事-記事インデックス間のリンク関係を整合的に管理することができる．

フィードバックエージェントは監視機能，およびフィードバック機能を持つ．どの記事が新聞社の記事として利用されるかを監視機能で把握し，その結果を随時通信社へフィードバックを行う．通信社-新聞社間での情報交換は記事インデックスを介して行う．記事インデックスに掲載状況のノードを追加，修正することで，通信社は掲載状況を把握できる．

5. まとめ

本稿では，ニュースの重要度を決める価値付けについて議論し，ニュースの価値観を様々な観点から決めるために掲載状況や閲覧状況を利用した．それらのメタ情報を収集するために，既存の記事配信フローを活かし，KDML を用いて拡張した記事インデックスを共有するシステムを構築した．

今後の課題は，掲載状況や閲覧状況を有効的に活かすシステムの実現である．そのために，過去のある期間における重要記事リストの作成において，記者による手作業とシステムによる自動生成とを比較検討する必要がある．

参考文献

[1] 児玉 政幸, 伊藤 正都, 大囿 忠親, 新谷 虎松, ”次世代記事編集システムにおける NewsML を用いたメタ情報配信について” 合同エージェントワークショップ&シンポジウム 2006 論文集, 2006.

[2] 大囿 忠親, 新谷 虎松, 青崎 保好, ”NewsML 記事のためのメタ記事記述言語 KDML について” 第 69 回情報処理学会全国大会論文集, 2007.

[3] 伊藤 正都, 児玉 政幸, 大囿 忠親, 新谷 虎松, 黒田 義和, 青崎 保好”メタ記事記述言語 KDML における記事リンク整合性管理機構の試作” 第 69 回情報処理学会全国大会論文集, 2007.

[4] 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” 情処学情報システムと社会環境論研報, Vol.2002, No.056, pp.1-8, 2002.

[5] 大川原雄也, 大囿忠親, 新谷虎松, ”言語モデルに基づく階層型クラスタリングを用いたトピック分析” 第 69 回情報処理学会全国大会論文集, 2007.

[6] Dolf Trieschnigg, Wessel Kraaij, ”TNO hierarchical topic detection report at TDT”, Topic Detection and Tracking 2004(TDT2004) Workshop, 2004.