社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－NL－180 (18)
2007／7／25

情報の信頼性分析に向けた評価データおよびプロトタイプシステムWISDOM

宮森 恒† 赤峯 享† 加藤 義清† 兼岩 憲† 角 薫†

乾 健太郎†,†† 黒橋 禎夫†,†††

† 情報通信研究機構 知識処理グループ 〒619-0289 京都府相楽郡精華町光台 3-5
†† 奈良先端科学技術大学院大学 情報科学研究科 〒630-0192 奈良県生駒市高山町 8916-5
††† 京都大学大学院 情報学研究科 〒606-8501 京都市左京区吉田本町
E-mail: †{miya,akamine,ykato,kaneiwa,kaoru}@nict.go.jp, ††inui@is.naist.jp, †††kuro@i.kyoto-u.ac.jp

あらまし 本稿では，情報の信頼性を自然言語処理に基づいて分析する際に必要となる評価用データおよびプロトタイプシステムWISDOMについて述べる．われわれは，ウェブ上のテキストを主な対象として，情報信頼性を分析することを目指したプロジェクトを2006年4月より進めている．本プロジェクトでは，ウェブ上の情報の信頼性を，情報内容，情報発信者，情報外観，社会的評価といった4つの基準で捉えることを提案しており，これらを述語項構造を単位とする自然言語処理によって論理的に分析・組織化することを目指している．本稿で述べる評価用データは，これら種々の分析処理の学習・検証用データとして構築されたものであり，時事問題，医療問題等の20トピックを選定し，各100ウェブページを収集して，各評価尺度のデータを人手で付与したものである．また，情報信頼性を多角的に評価するプロトタイプシステムWISDOMを開発した．本システムを用いて上記評価尺度で条件を様々に変化させて情報閲覧することにより，興味のトピックについて，信頼できる情報をより確実に見極めることができるようになる．
キーワード Web, 情報信頼性, 情報分析, 自然言語処理, 述語項構造, 評価尺度

Evaluation Data and Prototype System WISDOM for Information Credibility Analysis

Hisashi MIYAMORI†, Susumu AKAMINE†, Yoshikiyo KATO†,

Ken KANEIWA†, Kaoru SUMI†, Kentaro INUI†,††, and Sadao KUROHASHI†,†††

† Knowledge Clustered Group, NICT 3-5, Hikari-dai, Seika-cho, Souraku-gun, Kyoto, 619-0289 Japan
†† Graduate School of Information Science, Nara Institute of Science and Technology 8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192 Japan
††† Graduate School of Informatics, Kyoto University Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
E-mail: †{miya,akamine,ykato,kaneiwa,kaoru}@nict.go.jp, ††inui@is.naist.jp, †††kuro@i.kyoto-u.ac.jp

Abstract Evaluation data and a prototype system named WISDOM used for analyzing information credibility based on natural language processing are described. Our group started the Information Credibility Criteria project in April, 2006, mainly to analyze the credibility of information (text) on the Web. The project proposes to capture information credibility based on four criteria (content, sender, appearance, and social valuation) and aims to analyze and organize them logically using natural language processing based on predicate argument structure. The evaluation data were developed as learning and verifying data for these various analysis modules, and were composed of manually-annotated data based on each evaluation criteria about pre-selected 20 topics such as current events and medical issues with 100 pages per topic being collected from the Web . The prototype system WISDOM was developed to provide information credibility from different perspectives. Users will be able to find credible information more reliably by browsing information using different evaluation criteria and conditions provided by the system.
Key words Web, information credibility, information analysis, natural language processing, predicate argument structure. credibilitv criteria

1. まえがき

ネットワークや計算機端末の発達に伴い、さまざまな情報がウェブを通して利用できるようになった．知らない言葉や新しい話題を知るためにウェブを利用することはすっかり日常的な風景になった．

しかし，現在のウェブに存在する情報の量は膨大になったが，情報の質という意味では日常生活に本当に役立つ情報と何の根拠もない嘘やデマといった情報が玉石混交の状態で存在している．現在，数々の検索エンジンが利用可能であり，該当する情報の収集を行うことはできるが，一般利用者がこれら検索結果から情報の信頼性や信憑性の判断を行うのは極めて困難な状況にあるといわざるを得ない．

われわれは，既存の検索エンジンがもつ問題点の一つとして，検索結果が単一の評価軸で提示されることがあると考えている．例えば，アガリクスという健康食品について既存の検索エンジンを用いて調べてみると，健康によいという宣伝をするページが上位に表示され，その他の情報は下位のどこかに埋もれてしまい，本当に健康によいのかどうかを判断するのは極めて難しい．

信頼できる情報を見つけるためには，妥当な内容が書かれてるか，どんな人・組織が書いているか，連絡先や情報源は明示されているか，社会的にはどう見られているかといった多角的な観点から情報を整理し利用者に提示する必要があり，既存の検索エンジンとは異なる情報分析システムの構築が不可欠となる．

本稿では，このような情報分析システムを自然言語処理に基づいて実現する際に必要となる評価用データおよびプロトタイプシステム WISDOM について述べる．ここで述べる評価用データは，情報の信頼性を分析する際の種々の解析処理の学習・検証用データとして構築されたものであり，時事問題，医療問題等の 20 トピックを選定し，各 100 ウェブページを収集して，各評価尺度のデータを人手で付与したものである．また，WISDOM は，ウェブ上の情報信頼性を多角的に評価するプロトタイプシステムであり，本システムを用いて信頼性評価の各尺度で条件をさまざまに変化させて情報を閲覧することにより，利用者は自分の興味のあるトピックについて，信頼できる情報をより確実に見極めることができるようになる．

本稿の構成は以下の通りである．2 章でわれわれが進めている情報の信頼性分析プロジェクトの概要を説明し，3 章で評価用データの詳細について述べる．4 章でプロトタイプシステム WISDOM を紹介し、最後に 5 章でまとめる．

2. 情報の信頼性分析プロジェクトの概要

情報通信研究機構知識処理グループでは，2006 年 4 月より情報の信頼性分析プロジェクトを進めている．本プロジェクトでは，主にウェブ上のテキスト文書を対象として，信頼できる情報を検出・組織化し，背景的知識，事実，論点，意見分布などを的確に抽出する基盤技術の確立を目指している [1].

われわれは，ウェブ上の情報の信頼性を，情報内容，情報発信者，情報外観，社会的評価といった 4 つの基準で捉えることを提案しており，これらを述語項構造を単位とする自然言語処理によって論理的に分析・組織化することを目指している．

情報内容はウェブページの本文に書かれている内容に着目した観点で，文内容の信頼性，分類，要約といった処理が含まれる．情報発信者は，発信者の身元に着目した観点で，所属の分類やその分野での専門性の有無に関する処理が含まれる．情報外観は，ウェブページの見た目に着目した観点で，情報ソースや連絡先の明記，デザインや文体の適切さに関する処理が含まれる．社会的評価は，他の利用者がどのような見方をしているかに着目した観点で，意見表現の有無や意見分析に関する処理が含まれる．

また，述語項構造とはテキスト文書中の「誰が何をどうした」といった文中の単語間の意味的関係であり，これを単位とした分類，要約，意味解析や，既存知識との比較，整合性検証といった論理的分析を行うことで，信頼できる情報を的確に利用者に提示することができるようになると考えている．

3. 情報の信頼性分析に向けた評価データ

情報の信頼性分析に向けた評価データを構築した．本評価データは，信頼性分析システムの学習用および検証用データとして利用される．情報信頼性が特に問題となる 20 トピックを選定し (表 1)，各トピックについて様々な発信源から約 100 ウェブページを収集した．収集したウェブページに対し，情報内容，情報発信者，情報外観，社会的評価の各評価尺度で分析されるデータを人手で付与した．

表 1　選定されたトピック
Table 1　Selected topics

時事問題などの情報信頼性	医療問題，健康食品などの情報信頼性
(A1) 捕鯨問題	(B1) アガリクス
(A2) ＢＳＥ問題	(B2) がんの予防
(A3) 子供の体力低下	(B3) アンチエイジング
(A4) 少子化問題	(B4) メタボリックシンドローム
(A5) 年金制度	(B5) 携帯電話の電磁波対策
(A6) ＮＨＫ受信料問題	(B6) マイナスイオン
(A7) ＣＯ２削減問題	(B7) ダイエットのタブレット
(A8) パレスチナ問題	(B8) ダイエット食品
(A9) スーダンの民族紛争	(B9) 京都の観光
	(B10) 横浜の観光
	(B11) 京阪奈 [特定地域] での産婦人科

まず，情報内容の尺度からは，ページ分類のためのクラスタリング，概念分類のためのオントロジ，質問応答のための質問応答ペアに関するタグを付与した．さらに，各ページごとに重要文，文内容を総合判断した際の信頼性・有用性のタグを付与し，各文ごとに談話構造に関するタグを付与した．

情報発信者の尺度からは，各ページごとに，個人，営利・非営利団体，報道機関といった発信者の分類，各トピックごとの発信者の信頼度に関するタグを付与した．

情報外観の尺度からは，情報ソース・連絡先・専門性等が明

記されているか，デザイン，データ最新度，文体に関するタグを付与した．

社会的評価の尺度からは，各文ごとに，意見性を含む表現の有無，意見を含む場合はモダリティや意見保持者，意見対象，意見極性，意見分類といったタグを付与した．

以下，詳細を説明する．

3.1 クラスタリング

クラスタリングについては，収集されたウェブページ群を対象とし，これらを各トピックごとに5つ程度のクラスタに分割し，各クラスタに対してラベルを付与する．例えば，捕鯨に関するラベルとしては，

- 国際捕鯨委員会
- 捕鯨反対派の活動・意見
- 調査捕鯨
- 捕鯨の歴史に関する情報提供
- 捕鯨に関する感想・意見

といったラベルが付与される．このラベル付与は，原則として各トピックに対して1種類のみ行うが，顕著な複数観点がある場合は，最大3種類の観点からのラベル付与を行った．

また，クラスタリングを行う際には，

- 発信者属性や更新日時のような特定項目による分類でなく，ページの本文に基づい分類すること，
- クラスタを一覧することで人間がページ群全体の傾向を把握できること，
- 付与されたラベルと各ページの内容に乖離がないこと，
- 各ページは必ず一つのクラスタのみに属すること

といった点に留意した．

これらクラスタは各トピックごとに下記XML形式で作成される．

3.2 オントロジー

オントロジーは，各トピックを体系的に理解する際の基本知識として利用され，トピック別に基本概念が体系化された知識として機械利用可能な形式で表現される．

具体的には，収集されたウェブページ群を基本とし，必要に応じて他のページも参照しながら．トピック全体を表す語彙をルートとして，トピックの下位概念を表す語彙をノードとしたツリー形式で記述される．また，各ノードには，語彙の定義(一般的でない用語のみ)，及び，関連ウェブページへのリンクが必要に応じて付与される．

3.3 QA

QAは，各トピックについての質問と回答の集合である．ここでは，トピックごとに30程度の質問とその回答および回答が記載されたURLから構成されるデータを作成する．

質問は，そのトピックに関して頻出する質問をウェブページから抽出し，1文の疑問文で記述する(例．(a) 国際捕鯨委員会には，どんな国が加盟していますか？(b) ミンククジラがサンマやサケを捕食しているって本当ですか？)．

(a)のような誰でも単一の正しい回答が可能な質問については，単一の回答を示し，その回答を直接的に説明している信頼度の高い発信者のURLを記述する．また，(b)のように，人(組織)や状況により，回答が異なる質問については，回答の要約，及び，異なる意見を併記して，各意見が記載されたURLを記述する．

3.4 要　　約

要約は，収集された各ウェブページに対し，ユーザがそのページにアクセスすべきかどうかをひと目で簡単に把握できることを目的に作成され，タイトルと重要文の2種類から構成される．

タイトルについては，各ページのトピックについて記述された部分に基づき，アノテータが第三者の立場で30文字程度で付与した．例えば，「〜に関する〜のインタビュー記事」「〜に関する雑感」「〜に関するユーザ同士の意見交換」といったタイトルを与えた．

また，重要文については，各ページのトピックについて記述した部分から，内容や特徴を表現している重要文を1〜3文抽出し，内容が把握しやすいように整形して記述した．この際，連続した文になっていない場合は，文が連続していないことを示す「…」を文頭に記述した．

3.5 ページ信頼度

各ページごとに，本文の内容，情報発信者，情報外観の3つの観点から信頼性に関するタグを付与する．

まず，本文の内容については，ページの本文を読み，その内容が信頼できるかどうかを7段階で評価する(内容の信頼性)．この時，発信者が誰であるかや，ページのデザイン等を考慮せず，文章の内容だけから信頼性を判断する．また，ページの内容について，分析対象のトピック名に照らし合わせてどれだけ有用であるかを3段階で評価する(内容の有用性)．

情報発信者については，情報発信者の固有名と分類名を記述する．

情報発信者の固有名は，個人名，個人が属する団体名，情報発信サイトの団体名の3項目で表現される．これら項目のうち2項目以上特定できる場合，最も情報発信に関して責任を持つ項目は代表発信者として記述される．例えば，A学会のサイトにあるB株式会社C研究所のDさんの記事のページに対する情報発信者の固有名は，個人名=D，個人が属する具体的な団体名=B株式会社C研究所，情報発信サイトの団体名=A学会となり，D@C@B@Aで表される．また，代表情報発信者=Dとなる．

情報発信者の分類名は，表2のいずれかを記述する．

情報外観については，3の各項目が該当するかどうかについて記述する．

最後に，対象ページについて，情報の内容，発信者，外観を含む全ての要素を考慮した上での信頼性を7段階で評価した値を付与する．

3.6 談話構造

談話構造については，各文を対象として，文と文，および，文と節の間の関係を表すラベルが付与される．タグの種別は「原因」「条件」「目的」「方法」「逆接」「並列」「例示」の7種類とし，関係を示す部分の開始・終了文節番号，関係を受ける部分の開始・終了文節番号と合わせて記述される．

表 2　発信者の分類名
Table 2　Sender class

1. 個人	(ア) 有識者・専門家・著名人 (イ) 一般 (ウ) 匿名・ハンドルネームのみ
2. 団体	(ア) 営利団体 　企業 　業界団体 (イ) 非営利団体 　行政 　公益法人等 　大学 　学会 　任意団体 (ウ) 報道機関 　新聞社 　雑誌 　テレビ・ラジオ
3. 個人の集合	(ア) 参加者は実名 (イ) 参加者は匿名
4. その他	素性はわかるが 1-3 のいずれにも当てはまらない場合
5. 不明	個人の匿名というレベルとは異なり，個人か団体かすら，わからない場合

表 3　情報外観の評価項目
Table 3　Evaluation items for appearance

1. 情報ソース、参考文献の明記
2. メールアドレス，素性・居所の明記 - 組織の場合，住所，電話番号，組織概要 - 個人の場合，経歴，所属先団体・素性
3. 専門性能力の明記 - 専門機関／専門家であることの記載 - 専門知識・情報の引用・リンク - 個人の見解 - プライバシーポリシー有無 - 複数言語での提供
4. デザインの適切さ
5. ページの使いやすさ - 上位ページのわかりやすさ，文字化け有無
6. サイトのデータの最新度 (最終更新日)
7. 広告 - トピックとの関連性・区別 - 自動ポップアップ広告の有無 - 活動参加の呼びかけ - 独自グッズの販売 - プレゼント情報や応募記事 - リンク先に店・企業広告が多い
8. 誤植
9. 文体，言葉の使い方 - 常体/敬体，書き言葉/話し言葉 - 頻繁な記号や顔文字 - 故意に変な言葉や表記

「原因」は，実際に起こった出来事を対象として，ある状況，状態になる (なった) 理由，要因等を表現する (例．眠かったので家に帰った)．

「条件」は，もしそうだったらという仮定の内容を対象とし，ある状況，状態になる前提，必要な事柄等を表現する (例．健が来たら奈緒美も来る)．

「目的」は，実現しよう，到達しようとして目指す事柄を表現する (例．彼女に会いに行った)．

「方法」は，ある目的を達するためのやり方，手段を表現する (例．茶がらをまいてほうきで掃くことで，たたみはきれいに掃除できます)．

「逆説」は，ある条件に対して予期しない結果であること，あるいは，条件と結果の間に食い違いがあることを表現する (例．健が来たのに，奈緒美は驚かなかった)．

「並列」は，同じレベルの内容が 2 つ以上存在していること (対照，選択を含む) を表現する (例．健が来たのに，浩は来なかった)．

「例示」は，あるカテゴリに対して具体例を挙げている部分を表現する (例．次の番号まで電話を下さい．< 電話番号 >)．

3.7　意　　見

意見については，各文を対象として，ある事態を表す部分と，それに対する何らかの心的態度を表す部分（モダリティ）に付与する．例えば，「地震が起きたようだ」の場合，事態＝「地震が起きた」，モダリティ＝「ようだ」とする．事態を表す部分のタグは「主観」「非主観」の 2 種類し，モダリティに付与するタグは「意思」「判断」「当為」「疑問」「婉曲」の 5 種類とした．

主観/非主観は，必要に応じて前後の文脈も参考にしつつ，事態部分が評価・感情・態度の表明をしているかどうかでいずれかのタグを付与した．事態部分が主観の場合，その主観の保持者，主観の対象，主観の極性 (肯定・否定の度合いを 7 段階評価)，主観の分類，主観表現の入れ子の有無，トピックとの関連性といったタグも付与した．

モダリティは「文末表現に表れる筆者の心的態度」ととらえ，複数のモダリティが混在することを許して広めに取得した．「意志」は，これからやろうと思う何かについての態度を表し，意志，申し出，勧誘，命令，禁止，依頼，願望の表現を含む．「判断」は，話者の判断の確からしさを表し，推量，可能性，様態，伝聞の表現を含む．「当為」は，一般的常識に照らした事態の必要性，望ましさを表し，当為，許可の表現を含む．「疑問」は明らかな疑問文で，「婉曲」は，意思・判断・当為・疑問のいずれでもない心的態度の表現である．

3.8　データ形式

上記データは，ページ群に対するデータ，各ページに対するデータの 2 種類の XML データで記述される．

ページ群のデータには，クラスタリング，オントロジ，QA が含まれる．また，各ページのデータとして，要約，内容・発信者・外観によるページ信頼度，談話構造，意見が含まれる．各ページのデータは，KNP 等の構文解析と親和性を高めたウェブ文書用標準フォーマット [2] を拡張することで記述した (図 1)．

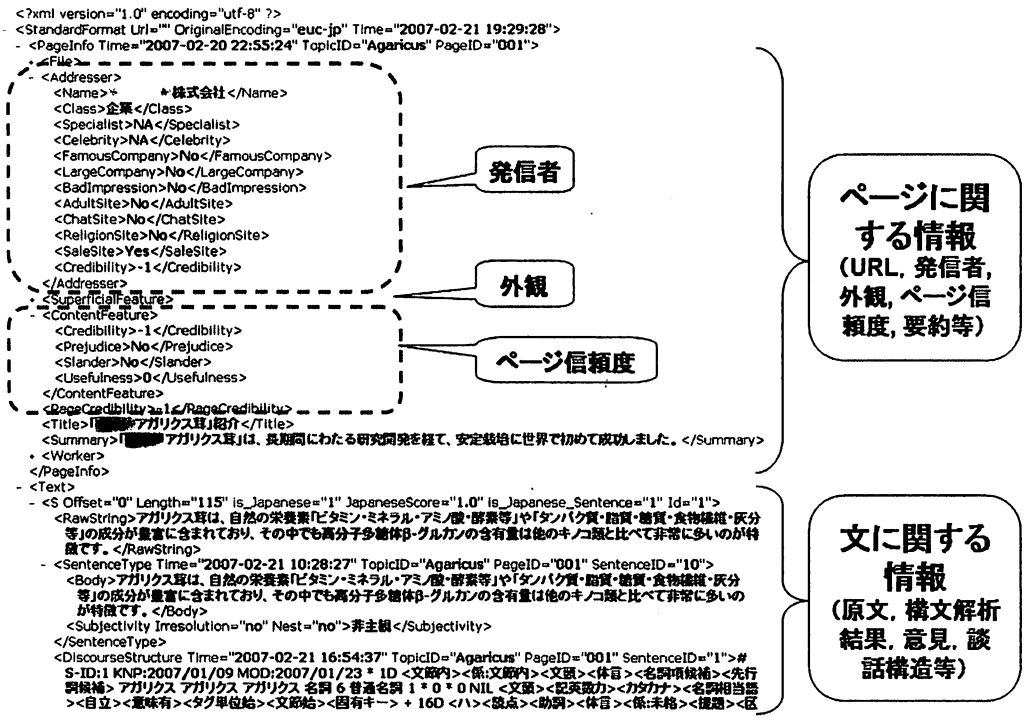

図 1　ウェブ文書標準フォーマットでの記述例

Fig. 1　Example description in standard format for Web documents

4. プロトタイプシステム WISDOM

情報信頼性を多角的に評価することを目的としたプロトタイプシステム WISDOM を開発した．本システムを用いて上記の各評価尺度で条件をさまざまに変化させて情報を閲覧することにより，自分の興味のあるトピックについて，信頼できる情報をより確実に見極めることができるようになる．

WISDOM の概要を図 2 に示す．まず，クローラを用いてウェブからウェブページを収集し，ローカルストレージに蓄積する (step 1)．次に，クラスタリングやオントロジ生成などページ群に対する分析と，発信者・外観・内容に関するページ信頼度や意見解析など各ページに対する分析を行う (step 2)．分析結果は，それぞれウェブ標準フォーマット等の XML データとして格納される．

ユーザはウェブブラウザから WISDOM にアクセスし，分析したいトピックをキーワード入力する．分析ボタンを押すことにより，そのトピックに関連したウェブページを，内容や発信者で分類した一覧，概念分類した一覧，意見分布の一覧として表示でき，各項目による全体的な動向や概要を把握できる (図 3)．また，特定ページの情報を表示させることで，そのページの発信者や外観の適正さ，意見内容と全体の中での位置づけ等を確認することができる (図 4)．このように，利用者は，興味のあるトピックについてさまざまな観点から分析結果を閲覧することができ，信頼できる情報をより確実に見極めることができるようになる．

5. ま と め

情報の信頼性を自然言語処理に基づいて分析する際に必要となる評価用データ構築とプロトタイプシステム WISDOM について述べた．

現在のシステムは，評価データに基づいて生成された XML データを参照することにより動作しており，今後，ページ収集や各種分析処理を自動化・高精度化していく必要がある．現在，日本語ページの収集[4] や発信者分類[5] の自動化を順次進めており，今後もモジュールやデータ更新を随時行っていく予定である．

文　献

[1] 黒橋禎夫: 構造的言語処理による情報分析研究, 言語処理学会第 13 回年次大会 (NLP2007) ワークショップ (W2), pp.17-18, 2007.

[2] 新里圭司, 橋本力, 河原大輔, 黒橋禎夫: 自然言語処理基盤としてのウェブ文書標準フォーマットの提案, 言語処理学会第 13 回年次大会 (NLP2007), E3-1, pp.602-605, 2007.

[3] KNP. http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

[4] ICC-Crawler. http://kc.nict.go.jp/icc/crawl-ja.html

[5] 加藤義清, 黒橋禎夫, 乾健太郎: Web 文書の情報発信者クラス分類, 言語処理学会第 13 回年次大会 (NLP2007), D4-7, pp.891-894, 2007.

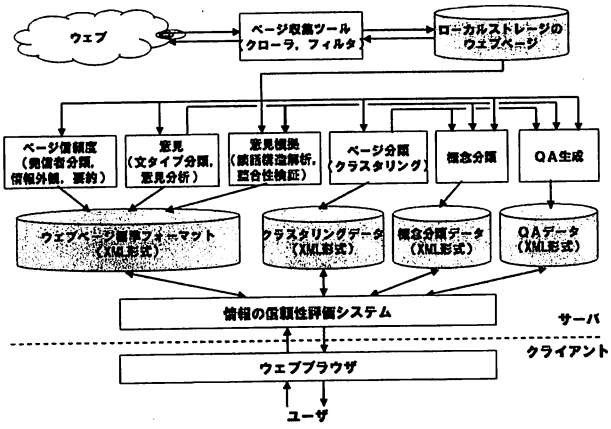

図 2　プロトタイプシステム WISDOM の概要

Fig. 2　Overview of prototype system WISDOM

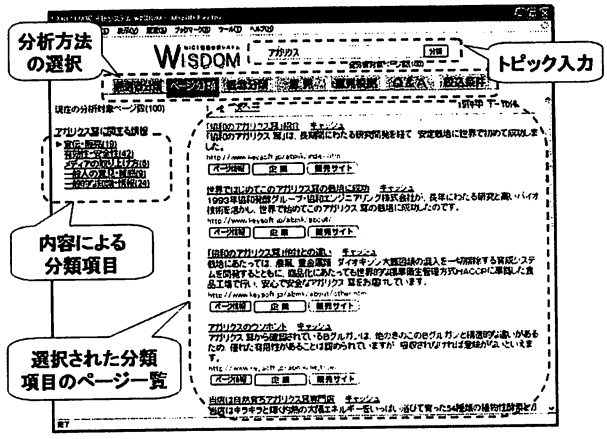

図 3　内容に基づくページ分類の例

Fig. 3　Example of classifying Web pages based on content

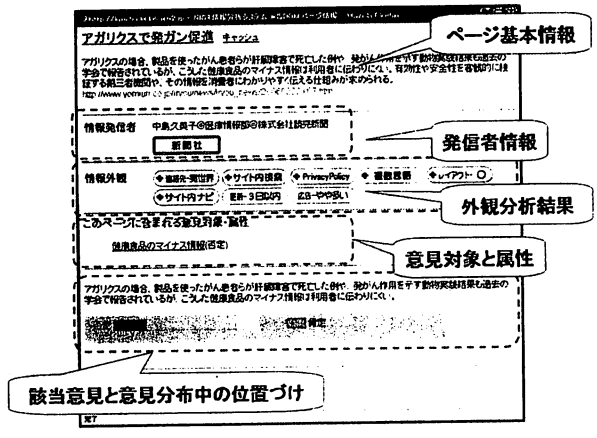

図 4　特定ページの分析例

Fig. 4　Example of analyzing a specific page