気になるキーワードから時事の話題を検索できるホットワードリンク™

HOTWORDLINK™ for Topical Word Extraction and Related Information Retrieval

岡本　昌之　■ OKAMOTO Masayuki
藤野　剛　■ FUJINO Go
根岸　伸一　■ NEGISHI Shinichi

近年，インターネット上では，旬（しゅん）の話題を抽出して可視化する技術が注目されている。東芝は，時事性の高い話題とその推移をキーワード群として表示し，話題の容易な把握とワンクリックで関連Webページを検索できる，話題抽出技術を用いたAVノートパソコン（PC）向け機能“ホットワードリンク™”を開発した。この機能は，2段階のクラスタリングによる話題抽出，統計処理による時事性の分析，話題のポジティブ／ネガティブ（P/N）判定，及び人名抽出を特徴としている。抽出された話題について被験者調査を行った結果，抽出されるキーワードと分析・生成される話題の推移グラフはユーザーにとって話題を把握しやすくさせることがわかった。

Technologies for extracting and visualizing topical news are a current trend in Web services. Toshiba has developed HOTWORDLINK™, a topic-extraction function for audiovisual-specialized notebook PCs. HOTWORDLINK™ visualizes topical news items and their trends and enables the easy retrieval of related Web pages with one click. The features of HOTWORDLINK™ include topic extraction with two-level clustering and statistical techniques, the classification of each topic into positive or negative, and person-name extraction. The results of experiments showed that the extracted topics and trend graphs contributed to the subjects' understanding.

1　まえがき

近年，Web上では，世間で旬（しゅん）の話題を表すキーワード（ホットワード）を抽出して可視化する技術やサービスが注目されている。例えば，Web検索履歴やWeblog（ブログ）の記事に含まれる頻出キーワードのランキング表示や，それらキーワードごとの頻度の推移をグラフ表示するサービスがある。しかし，キーワードが同時に出現する単語と雑多に提示されたり，グラフ形状と話題の変化との対応付けができないため，ある話題がどのような経過をたどったかを知るには，ユーザーが試行錯誤して調べる必要がある場合が多い。

東芝は，AVノートPCでの映像視聴の合間に，テレビや新聞で目にする旬な話題を経緯と併せて把握し，興味ある話題があればすぐに関連するWebページに到達することを支援する，“ホットワードリンク™”機能を開発した。この機能は，2006年8月に発売された当社AVノートPC Qosmioシリーズの新機能として製品化された[(1)]。ホットワードリンク™利用により，日々注目されている話題をキーワードとして把握するとともに，キーワードのクリックで関連Webページの検索まで行うことができる。

ホットワードリンク™には，当社のコア技術の一つである時事キーワード抽出技術が利用されている。2007年7月には，更に話題推移抽出，話題のポジティブ／ネガティブ（P/N）判定，タレント抽出，及び関連番組抽出の各機能を加えてサービスをリニューアルした。ここでは，これら機能の概要と使用される話題抽出技術について述べる。

2　ホットワードリンク™の概要

ホットワードリンク™は，Qosmioシリーズに搭載されているAV統合ソフトウェア“Qosmio AV Center”の機能の一つである。この機能では，時節の単語や人名，政治，社会，ビジネス用語などのキーワードがランキング表示される。インターネット環境への接続設定や，地域及び電子番組表の設定を行うと，毎日番組表を受信するタイミングでホットワードリンク™のトップページが当社のサーバからダウンロードされ，次の手順で利用できる。Qosmio AV Centerを起動してホットワードリンク™のボタンを押すと，Webブラウザが開きその日のホットワード一覧が表示される（**図1**）。ホットワードリンク™機能は，“ホーム”，“全画面表示”，及び“ながら見”のどのモードからでも呼び出すことができる。

ホットワードリンク™は，以下の4種類のキーワード群から構成される。

(1)　最近注目のホットワード　　直前の数日間に話題となっているキーワード
(2)　ずっと気になるホットワード　　最近注目のホットワードより長い期間（1週間以上）に話題となっているキーワード
(3)　急上昇ホットワード　　最近注目のホットワードのうち，特に取り上げられることが急増したキーワード

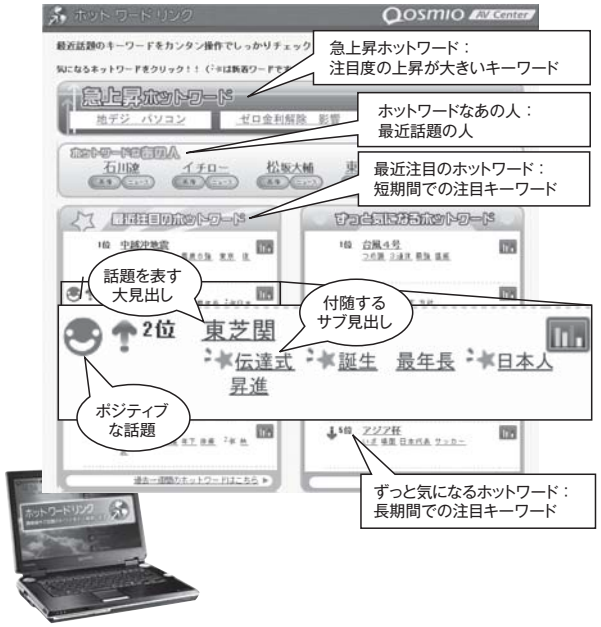

図1. ホットワードリンク$_{TM}$画面 — Webページに3種類の話題キーワード群と，話題に挙げられた人名が表示される。

Example of HOTWORDLINK$_{TM}$ display

(4) ホットワードなあの人　　最近注目されているタレントやスポーツ選手，政治家などの有名人

このうち，(1)と(2)は大見出しとサブ見出しで構成され，大見出しは話題を示す代表的なキーワードであり，サブ見出しは大見出しと関連性が強いキーワードである。星印が付いたキーワードは新着キーワードを示し，矢印の向きは前日からの順位の変化を示す。楽しい話題や明るい話題の場合には笑顔のアイコンが表示される。

また，それぞれの話題の大見出しの横にあるグラフのアイコンをクリックすると，話題の推移を知ることができる。グラフの縦軸は話題の注目度，横軸は日付を示し，話題の節目となるイベントがグラフに重畳して表示されるとともに，大見出しのキーワードに関連する番組が検索され表示される（**図2**）。

これらのキーワードは，毎日自動的に集計され更新されて注目度に応じたランク付けが行われ，注目度の推移に応じて分類され表示される。表示されたキーワードはWeb検索ページへのハイパーリンクとなっており，ユーザーは興味ある話題を表すキーワードをクリックすることで，その話題について調べることができる。

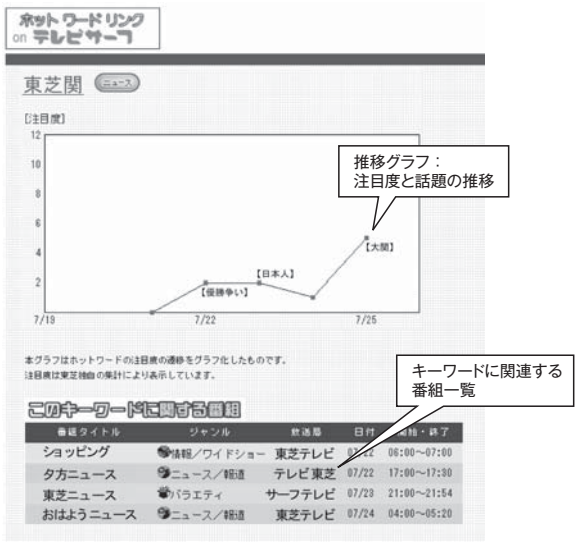

図2. 話題の推移 — 選択した話題にどのようなイベントが含まれ，注目度がどのように変化したかを推移グラフで表示する。また，関連番組一覧が表示される。

Trend of topic

3　話題抽出の流れ

ホットワードリンク$_{TM}$における話題抽出の概略を**図3**に示し，詳細を以下に述べる。

3.1　話題推移を考慮したクラスタリングとキーワード抽出

日々のニュースなどのテキスト情報を入力し，形態素解析（文章を一つ一つの単語に分割する処理）結果に基づき単語ベクトルを生成する。例えば，“東芝関が優勝争いに加わった”という文章は“東芝／関／が／優勝／争い／に／加わっ／た”という単語ベクトルに分けられる。

次に，クラスタリング処理によりそれぞれのテキストを話題に分類する。クラスタリングとは，文書処理で複数の文書を類似する文書グループに分類する技術である。その手法は，各文書をボトムアップにまとめていく階層的クラスタリングと，全体の集合をトップダウンに分割する非階層的クラスタリングとに大別できる。当社は，計算時間はかかるが比較的性質の良い分類結果を得やすい，階層的クラスタリングを選択した。処理の流れを**図4**に示す。

分類された各話題に対しては，C-value[(2)]という指標に基づいて複合語を含む代表キーワードが生成され，大見出しとして利用される。例えば，先の文書を含む話題からは“東芝関”というキーワードが大見出しとして抽出される。

次に，各話題に対して更に階層化クラスタリングを行うことで，各話題に含まれる詳細なイベントが抽出される。この処理では日時情報を重み付けとして利用することで，時系列的にイベントの推移が抽出される。これら各イベントの代表キーワードがサブ見出しとして利用される。例えば，優勝争いを演じ，その後大関に昇進したという流れであれば，優勝や大関，伝達式などのイベントを示すサブ見出しが抽出される。そのほかにも，効果的なキーワード抽出を行うために，不要語の除去や敬称表現の集約などの処理が行われる。

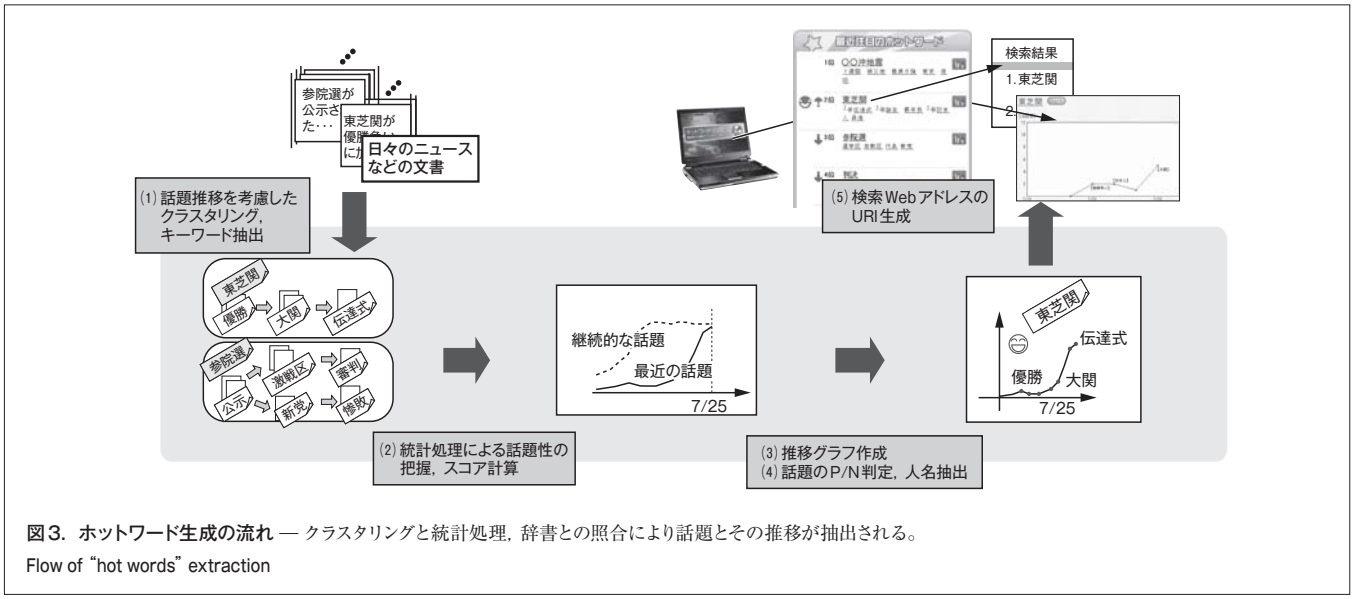

図3. ホットワード生成の流れ — クラスタリングと統計処理，辞書との照合により話題とその推移が抽出される。
Flow of "hot words" extraction

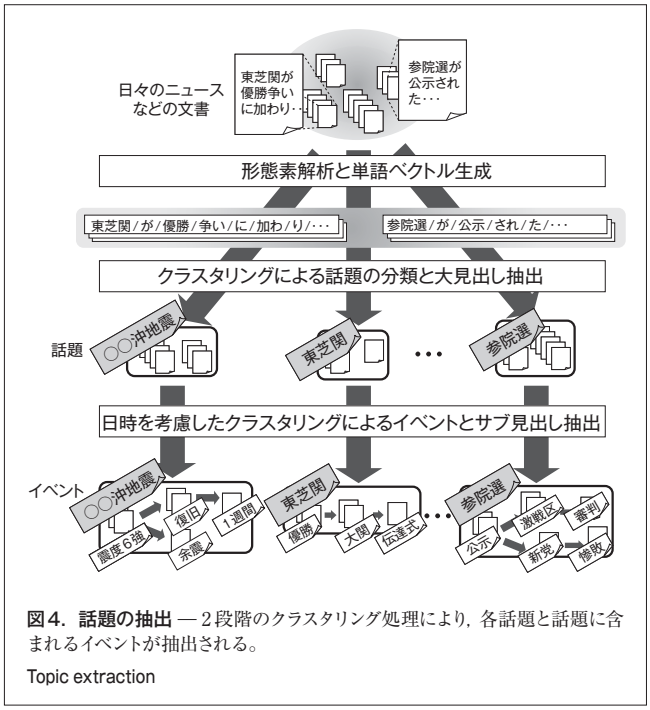

図4. 話題の抽出 — 2段階のクラスタリング処理により，各話題と話題に含まれるイベントが抽出される。
Topic extraction

3.2 統計処理による話題のスコア算出

3.1節の処理で得られた各話題に対し，話題性（話題の大きさ）や時事性（どれだけホットな話題か）を用いたランキングのためのスコア算出が行われる。

話題性については話題に含まれる文書数を用い，時事性については最近の日時に含まれる文書数が集計期間の文書数に対してどれだけ多いかという指標が用いられる。また，2章で述べた(1)の最近注目のホットワード（短期）と(2)のずっと気になるホットワード（長期）では，同じ処理が集計期間を変えて実行される。これにより，ここ数日で盛り上がっている話題と，しばらく継続している話題の両方が提示される。

3.3 話題キーワードの分類と推移グラフの生成

その後，長期の見出しから短期の見出しと重複して出現するものを除去し，短期の見出しのうち前日からスコアが大きく上昇したものを(3)の急上昇ホットワードとして取り出す。この処理により同じ話題が重複せず，また，異なる時間軸に応じたキーワード表示が行われる。また，話題クラスタ及びサブ話題クラスタの構造に基づき，各話題の推移グラフが生成される。サブ見出しは，推移グラフ上でその話題が出始めた日に重畳して表示される。したがって，図2に示すような推移グラフを見るだけで，ある話題の中でいつ，どのようなイベントが起きたか知ることができる。

3.4 話題のP/N判定と人名抽出

各話題に対してP/N判定を行う。話題を構成する文書に対し，当社が保有するP辞書とN辞書でマッチングを行い，それぞれのスコアに基づき算出された結果がPと判定されると笑顔のアイコンが付与される。同様にして，(4)のホットワードなあの人の名前も人名辞書のマッチング結果に基づいて抽出される。これらのマッチングにおいては，ほかのキーワードの部分文字列が誤って一致しないよう，形態素の境界判定が用いられる。

3.5 検索用URIの生成

各見出しに対応したワンクリックでのWeb検索を実現するため，検索キーワードとWeb検索エンジンのアドレスを表現するURI（Uniform Resource Identifier）が生成される。大見出しにはそのキーワードが，サブ見出しには"見出し AND サブ見出し"の検索キーワードが生成される。また，長いキーワードはそのままでは検索結果が出力されない可能性があるため，検索キーワード生成時に再度形態素に分割される。ただし，過度に分割されると適切な検索結果が得られないため，接頭語や接尾語などは連結される。

例えば，見出し“東芝関復活”とサブ見出し“優勝”からは，“東芝関 AND 復活 AND 優勝”という検索キーワードが生成される。また，人名の検索では，画像やニュースを効果的に検索するため，“画像”や“ニュース”などのキーワードで拡張された検索キーワードが用いられる。

4　キーワードからの話題把握に関する評価

当社開発の話題抽出エンジンにより抽出されたキーワードが，ユーザーに話題を想起させ，検索により関連するWebページに到達させられるかを調査した。対象は2007年7月23日～8月8日の17日間における短期及び長期の両期間で，上位の話題340件（短期，長期それぞれ170件）について，6人の被験者が1日当たり3～4人ずつ出力結果を確認した（**図5**）。

まず，キーワード群から時事の話題を想起できるか調べた結果，全体として“時事の話題を想起できる”が84.7％，“知らないが話題であるとわかる”が11.9％と，合計96.6％はユーザーが理解可能な程度にまとまっていた（図5(a)）。話題に含まれる個別のキーワードの有効性に関しても，94.9％のキーワードは適切であることがわかった。

次に，それぞれの話題の推移グラフを見たときに，話題の推移の状況がわかるかを調査した。その結果，全体として67.4％の話題については推移まで把握でき，特にイベントが4件以上の話題142件に関しては71.7％が推移まで把握でき，推移がわからない場合も含めて96.5％は把握できることがわかった（図5(b)）。

また，それぞれの大見出し340個とサブ見出し1,663個について，当日の11時に取得した10位以内の検索結果に適切な話題が含まれるか調査した。その結果，大見出しだけで検索する場合は78.8％，サブ見出しを含めた検索では88.8％と，サブ見出しを用いることでその話題の特定に効果があることがわかった（図5(c)）。これは，時事の話題を想起できる割合が84.7％であったことを考えると，サブ見出しを含めた場合には約4.1％は意外性のある結果を提示したと考えることもできる。

以上の結果から，ホットワードリンク$_{TM}$による話題提示は，キーワードや推移グラフの提示により，時事の話題やその推移の理解に寄与していると言える。

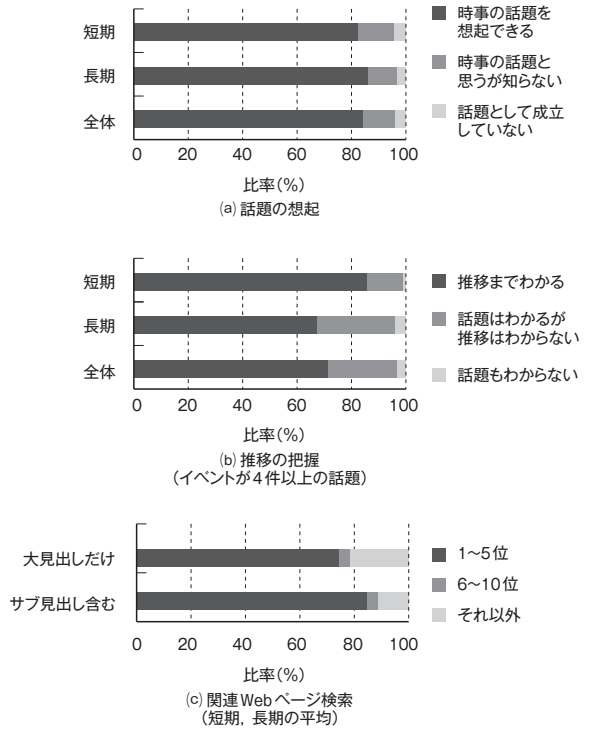

図5．話題抽出の効果 — 抽出された話題の96.6％はユーザーが把握でき，4件以上のイベントを含む話題の71.7％に対しては話題の推移がわかった。サブ見出しを用いた検索では，88.8％が10位以内に関連する話題を見つけられた。

Effect of topic extraction

5　あとがき

ここでは，Qosmioに搭載されたホットワードリンク$_{TM}$機能の概要と，利用される話題抽出技術について紹介するとともに，話題抽出の効果について被験者による評価結果を述べた。この評価により，キーワードと推移グラフによる簡潔な表示が話題把握に貢献することがわかった。

今後は，話題抽出の精度向上に取り組むとともに，様々なユーザーに適した話題提供の方式について検討を進めていく。

文　献

(1) (株) 東芝iバリュー クリエーション事業部．“AVノートPCに適した検索サービス「ホットワードリンク」の開始について”．<http://www.ivc.toshiba.co.jp/ivc/news/news20060807.html>，(参照2007-09-27)．

(2) Frantzi, K. ; Ananiadou, S. “Extracting nested collocations”. Proc. COLING-96. Copenhagen, 1996-08, ICCL. 1996, p.41－46.

岡本　昌之　OKAMOTO Masayuki, Ph.D.
研究開発センター　知識メディアラボラトリー研究主務，博士(情報学)。コンテクストアウェア技術及び話題抽出技術の研究・開発に従事。情報処理学会，人工知能学会，ACM会員。
Knowledge Media Lab.

藤野　　剛　FUJINO Go
ネットワークサービス事業統括部　iバリュー クリエーション事業部参事。デジタル機器向けネットワークサービス開発に従事。
iValue Creation Div.

根岸　伸一　NEGISHI Shinichi
PC＆ネットワーク社 PC第一事業部 PCマーケティング部。国内向けAVパソコンの商品企画業務に従事。
Personal Computer Div. -Japan & Asia Operations