番組視聴中に気になったシーンを簡単に検索できるキーワード抽出技術

Keyword Extraction for TV Program Information Retrieval without Interruption of PC Work

山崎　智弘　■ YAMASAKI Tomohiro
鵜沢　秀章　■ UZAWA Hideaki
林　英史　■ HAYASHI Eiji

近年，パソコン（PC）のAV機器化が急速に進んでいる。東芝は，AV-PCならではの効果的な番組視聴機能を実現するため，番組視聴中に気になったシーンについて，ワンクリックで簡単にインターネット検索できる機能として“気になるリンク™”を開発した。

気になるリンク™はキーワード抽出エンジンによって，ボタンが押された時点でのユーザーの検索欲求を推定し，キーワードの候補，及びキーワードの意味に応じた検索方法の候補を自動的に提示することを特長としている。このエンジンによって抽出されたキーワードの精度を評価した結果，高い再現率と適合率が得られ，気になるリンク™の有効性を確かめることができた。

PCs are rapidly joining the ranks of audiovisual (AV) equipment, a trend that is true for both notebook and desktop models.

Toshiba has developed a function called KININARU Link to enhance the experience of TV watching by AV-PC. By means of this function, users can easily retrieve information on a scene of interest when watching TV, with a single click. The features of Kininaru Link include estimation of the user's retrieval purpose when the button is pushed, and automatic enumeration of keywords and search methods suitable for the meaning of each keyword.

We evaluated the accuracy of keywords extracted by our engine and confirmed the efficiency of Kininaru Link through the high precision and high recall rate of the results obtained.

1　まえがき

地上デジタル放送の普及に伴い，PCのAV機器化が急速に進んでいる。しかし，現状はテレビ（TV）番組をPCで視聴できるようになっただけのことが多く，AV-PCならではの効果的な番組視聴機能が実現されているわけではない。例えば，番組視聴中に気になった“このお店のこの料理”についてもっと詳しい情報を調べたいと思っても，わざわざWebブラウザをTVアプリケーションとは別に起動してインターネット検索サイトを表示し，気になった話題を表すキーワードを入力し，検索結果の中から適切なページを吟味する，といった非常に煩わしい手順がいまだに必要である。

東芝は，このような煩わしさをなくすことを目的に，AV-PCならではの効果的な番組視聴機能として“気になるリンク™”を開発した。気になるリンク™は，番組視聴中に気になったキーワードについて，ユーザーが簡単にインターネット検索できるようにする機能である。ボタンが押された時点でのユーザーの検索欲求を推定し，キーワードの候補，及びキーワードの意味に応じた検索方法の候補を自動的に提示することが大きな特長である。

ユーザーは，提示された中から適切なキーワードと検索方法を選択するだけで簡単かつ的確に目的の情報が得られるため，PCで作業をしながらTVを視聴している場合でも，作業を中断することなく詳しい情報を気軽に調べることができる。

ここでは，気になるリンク™の概要，並びに気になるリンク™が使用しているキーワード抽出エンジンの処理の流れについて述べる。

2　気になるリンク™の概要

気になるリンク™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 に組み込まれている機能の一つである。この機能は，視聴中の番組に登場する人名，地名，及び店名などのキーワードを抽出し，一覧として表示する。利用方法の概要は以下のとおりである。

まず初期設定として，地上デジタル放送の視聴設定とインターネット環境の接続設定を行う。

その後，Qosmio AV Centerを起動してTV番組を視聴する。番組視聴中に，例えば次のような事がらについて“もっと詳しい情報をインターネット検索したい”と思った時点で，気になるリンク™ボタンを押す。

(1)　グルメ番組で紹介された店
(2)　ニュース番組の気になるニュース
(3)　旅行番組の宿泊や観光情報
(4)　音楽番組に出演したアーティスト

すると**図1**に示すようなキーワードリストの画面が開き，ボタ

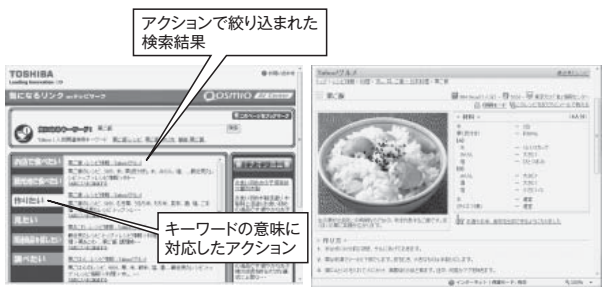
図1. キーワードリストの画面 ― 気になるリンク$_{TM}$ボタンを押したシーンに関連したキーワードの一覧と，キーワードの意味に対応したアクションの一覧が表示される。

Example of keyword list display

ンを押したシーンに関連したキーワードの一覧が表示される。キーワードは，そのシーンでの重要度の順に並べられている。

キーワードリスト画面では，“栗ご飯”（くりごはん）に対して“料理”，“タケノコ”に対して“食べ物”のように，それぞれのキーワードの意味を表すアイコンが表示されている。また，栗ご飯に対して“お店で食べたい”，“観光地で食べたい”，“作りたい”，“見たい”のように，それぞれのキーワードの意味に対応した検索方法を表すアクションの一覧が表示されている。

この画面で気になったキーワードのアクションをクリックすると，Webブラウザが起動し，検索結果一覧の画面が表示される。栗ご飯に対して“作りたい”のアクションをクリックした画面を**図2**(a)に示す。一般的なインターネット検索では検索結果の中から適切なページを吟味する必要があるが，気になるリンク$_{TM}$ではアクションによって絞り込まれた検索が行なわれているため，図2(b)に示すように，目的とする情報が簡単かつ的確に得られる。



(a) アクションをクリックした画面　　(b) 目的の情報画面

図2. 検索結果の表示画面例 ― アクションによって絞り込まれたインターネット検索が行われるため，簡単かつ的確に目的とする情報が得られる。

Example of search results display

表1. 意味情報とアクションの一覧

List of semantic attributes and search methods

意味情報（大分類）		アクション
(1) 動植物	(10) 地　位	(1) お店で食べたい
(2) 食べ物	(11) 連絡先	(2) お店で飲みたい
(3) 料　理	(12) 交　通	(3) 浸（つ）かりたい
(4) 組　織	(13) 無生物	(4) 観光地で食べたい
(5) 企　業	(14) 無形物	(5) 観光地で飲みたい
(6) チーム	(15) 温　泉	(6) 観光したい
(7) 地　域	(16) 飲み物	(7) 旅行したい（国内）
(8) 人　物	(17) 健　康	(8) 旅行したい（海外）
(9) キャラ	(18) その他	(9) 作りたい
		(10) 見たい
		(11) 動画を見たい
		(12) 株価を知りたい
		(13) 関連商品を探したい
		(14) 調べたい

気になるリンク$_{TM}$では**表1**に示すように，大分類で18種類の意味情報と14種類のアクションを用意している。キーワードごとにすべてのアクションを提示すると，アクションを選択する作業が煩雑になるだけでなく，意味によっては適切でない組合せが提示されてしまう可能性がある。そこで，内部的に意味情報とアクションの対応表を持ち，常に少数の候補の中から選択できるようにした。

3　キーワード抽出の流れ

気になるリンク$_{TM}$では，番組視聴中にシーンに関連したキーワードを抽出するため，番組全体の説明テキストと各シーンの説明テキストを入力として用いる。説明テキストは，十分な情報量を持っていればどのようなものでも入力として用いることができる。キーワード抽出処理は情報源が何であるかには特に依存しないため，以下では，番組全体の説明テキストと各シーンの説明テキストが与えられたものとして処理の流れを説

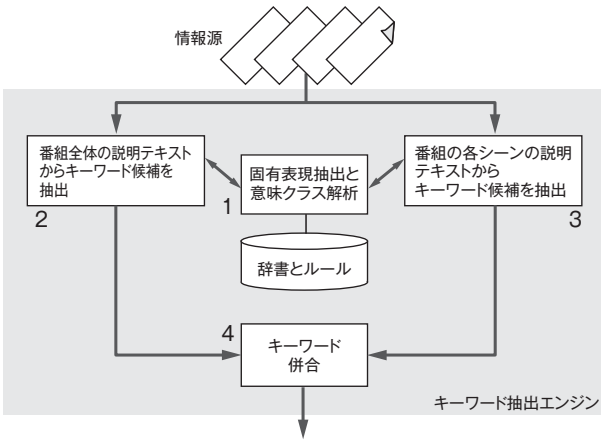


図3. キーワード抽出エンジンの構成 ― キーワード抽出エンジンは，固有表現抽出と意味クラス解析を中心に，番組全体の説明テキストと各シーンの説明テキストからキーワードの抽出を行う。

Configuration of keyword extraction engine

一般論文

明する。

図3に示すように，キーワード抽出エンジンは大きく分けて四つのモジュールから構成される。

一つ目のモジュールは，キーワードへの意味情報の付加を行うため，固有表現の抽出と意味クラスの解析[(1)]を行う。二つ目のモジュールは番組全体の説明テキストから，三つ目のモジュールは番組の各シーンの説明テキストから，固有表現抽出を用いてシーンに関連したキーワード候補の抽出を行う。四つ目のモジュールは，番組全体の説明テキストと各シーンの説明テキストから抽出されたキーワード候補を併合し，キーワードの意味情報に応じて検索方法の一覧を作成する。詳細を以下に述べる。

3.1　固有表現抽出と意味クラス解析

一つ目のモジュールは，テキストを形態素解析した結果を基にキーワードへの意味情報の付加を行う。意味情報は，人名や地名のような固有名詞のクラスと，日時や金額のような定型表現のクラスなど100種類以上あり，辞書には約11万個の語いが登録されている。また，辞書だけではなくルールを用いた意味情報の推定も行う。例えば“いさきの塩焼き”の場合，“いさき（動物）”，“塩（物質）”，及び“焼き（調理法）”と，“動物＋物質＋調理法→料理”というルールから，料理という意味情報が推定される。ルールは，自立語と付属語から成る文節間の関連を記述しており，約500種のルールが登録されている。

固有表現抽出と意味クラス解析の結果，“みそラーメン”に対する“ラーメン”のように，より長いキーワードの部分文字列として含まれるものが同時に抽出されることがある。より長いキーワードのほうが具体的で，ユーザーの検索欲求により合致すると考えられるため，このモジュールでは，部分文字列として含まれるものは抽出結果として採用しない。

3.2　番組全体の説明テキストと各シーンの説明テキストからのキーワード候補抽出

二つ目のモジュールは，番組全体の説明テキストに対して解析を行うことでキーワード候補を抽出する。こちらは視聴中の番組の主題を表すキーワードを抽出することが目的であり，ユーザーがボタンを押したシーンに特に関連していなくてもよい。

三つ目のモジュールは，番組の各シーンの説明テキストに対して解析を行うことでキーワード候補を抽出する。番組全体の説明テキストと異なり，各シーンの説明テキストからは，ユーザーがボタンを押したシーンに関連したキーワードを抽出することが要求される。そのためこのモジュールでは，番組内の話題の切替わりを考慮し，ボタンを押した時刻の120秒前から5秒後までの説明テキストに対して解析を行うことでキーワード候補を抽出するものとした。

120秒前から5秒後までという範囲は，次のようなトレードオフを考慮して設定したものである。

(1)　広いと話題が複数にまたがるため，シーンに関連していない前の話題のキーワードまで抽出してしまう。

(2)　狭いと入力としての説明テキストが少なくなるため，ユーザーの検索欲求に合致するキーワードを抽出できない。

120秒前から5秒後までの各シーンの説明テキストから抽出されたキーワード候補は，ボタンを押した時刻に近いほどユーザーの検索欲求に合致すると考えられるため，それぞれのキーワード候補には時刻の差に応じた重要度がベースとして与えられる。最終的な重要度は，かっこなどで強調されている場合は上げる，段落の切れ目では前の段落のものを下げるなど，いくつかのヒューリスティック[(注1)]に基づいて決定される。

3.3　キーワード併合

四つ目のモジュールのキーワード併合部は，番組全体の説明テキストから抽出したキーワード候補に意味情報に応じた重要度を与え，二つのパスによるキーワード候補を合わせて重要度の高いものから順に選択していくことで，最終的なキーワード候補を出力する。このとき，ストップワード[(注2)]辞書に含まれる検索キーワードとしてふさわしくないものは削除される。

一つのキーワードに対して複数の意味情報が付加されるときは，まず健康，次に無生物，地位，キャラクター，人物，飲み物，料理，食べ物，動植物，チーム，組織，企業，温泉，地域，交通，無形物，そして連絡先という優先順に意味情報が決定される。アクションはすべての意味情報に対応するものが併合される。ただし人物については，適切でないアクションが表示される可能性があるため，アクションの併合を行わないものとした。

4　キーワード抽出エンジンの評価

この機能の有効性を検証するため，キーワード抽出エンジンによって抽出されたキーワードの数，並びに再現率と適合率を評価した。評価の対象期間は，2007年09月20日から2007年10月5日までの2週間である。

初めに，抽出することができたキーワードの数が日時や時間帯などによってどの程度異なるかを調査した。抽出することができたキーワードの1時間当たりの平均数を，月曜日から日曜日までの00：00～23：59，月曜日から日曜日までの19：00～23：59，及び土・日曜日の00：00～23：59のそれぞれについて，主要放送局ごとに算出した結果を**図4**に示す。月曜日から日曜日までの19：00～23：59においては，各放送局とも非常にたくさんのキーワードが抽出されている。また，どの放送局でも，月曜日から日曜日までの00：00～23：59において，平均で1時間当たり20～30個のキーワードが抽出されている。このことから，よく視聴する放送局や時間帯はユーザーごとに

(注1)　試行錯誤的に発見した手法。
(注2)　あまりにたくさん検索にかかりすぎるので，検索精度向上のために検索対象から除外せざるをえないことば。

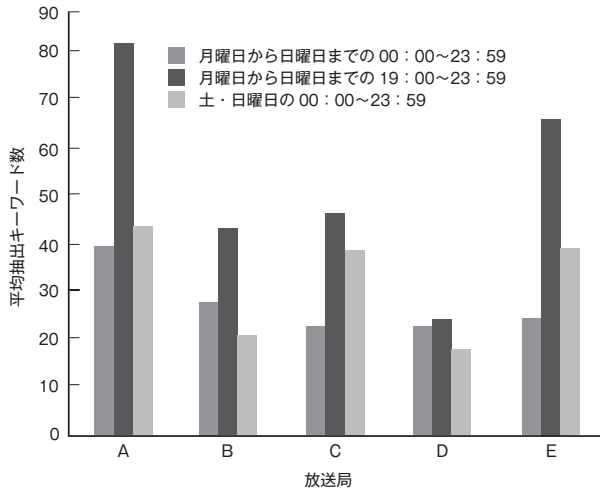


図4. 1時間当たりの平均抽出キーワード数 — A～Eのどの放送局も，月曜日から日曜日までの00：00 ～ 23：59の平均では20 ～ 30個，月曜日から日曜日までの19：00 ～ 23：59の平均では50個程度，土・日曜日の00：00 ～ 23：59の平均では20 ～ 40個のキーワードが抽出されている。

Average number of extracted keywords per hour

異なるものの，平均的には十分な量のキーワード一覧を提示できることがわかる。

次に日時，時間帯，及びジャンルを分散させた40番組を選択し，抽出されたキーワードの再現率と適合率を評価した。評価に用いる正解及び不正解のデータは，それぞれの番組で抽出されてほしい及びほしくないキーワードを被験者6人（1番組当たり2人）が確認することによって作成した。

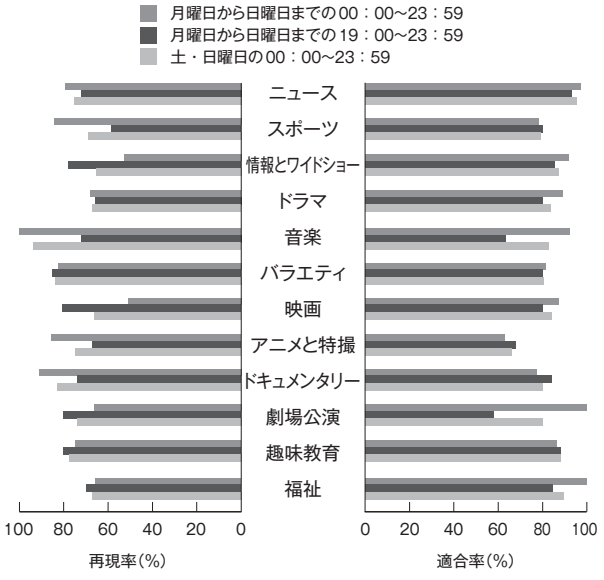


図5. 番組ジャンルごとの再現率と適合率 — 音楽，バラエティ，及びドキュメンタリーでは80％を超える高い再現率が得られ，アニメと特撮以外では80％を超える高い適合率が得られた。

Precision and recall of each genre

再現率については，正解及び不正解データを作成するときのキーワードの切出し範囲が被験者によってばらつくため，正解データと部分一致するキーワードに基づく評価を行った。**図5**に示すように，気になるボタンがあまり押されないと思われる映画や福祉では60％ほどにとどまっているが，音楽，バラエティ，及びドキュメンタリーでは80％を超える高い再現率が得られている。

適合率については，ユーザーの興味を引き，検索したいと思わせるキーワードを提示することも重要であると考えられるため，不正解以外のキーワードに基づく評価を行った。図5に示すように，説明テキストが十分な情報量を持っていないと思われるアニメと特撮は60％ほどにとどまっているが，全体的に80％を超える高い適合率が得られている。

5　あとがき

ここでは，気になるリンク$_{TM}$機能の概要と，この機能で使用しているキーワード抽出エンジンの処理の流れについて述べた。また，このエンジンによる抽出キーワード数，並びに再現率と適合率を評価し，この機能の有効性を検証した。

今後は，更なるキーワード抽出の精度向上に取り組むとともに，キーワードの抽出方式や検索方式をユーザーごとに最適にするため，履歴の活用を検討していく。

文　献

(1) 市村由美，ほか．“質問応答と日本語固有表現抽出および固有表現体系の関係についての考察”．自然言語処理研究会報告．東京電機大学，2004-05，情報処理学会．東京，2004，p.17－24.

山崎　智弘　YAMASAKI Tomohiro
研究開発センター　知識メディアラボラトリー。
テキストからのキーワード抽出及び話題抽出の研究に従事。
Knowledge Media Lab.

鵜沢　秀章　UZAWA Hideaki
ＰＣ＆ネットワーク社　ＰＣ開発センター　ＰＣソフトウェア設計第二部。ＰＣソフトウェアの開発に従事。
PC Development Center.

林　　英史　HAYASHI Eiji
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。デジタル機器向けネットワークサービスの開発に従事。
iValue Creation Div.

一般論文