5D-5

ＥＰＧメタ情報からの注目キーワード抽出手法の検討と評価

山下　道生†

株式会社　東芝　プラットフォーム＆ソリューション開発センター†

1 はじめに

近年のテレビやレコーダの高性能化と、ハードディスクの大容量化により、個人でも複数チャンネルの全番組録画が可能となった。録画予約操作なしで番組を丸ごと録画し、好きな時に見たい番組が視聴できるようになったことで、膨大な録画番組の中から自分の好みに合う番組を探し出す機能の要望が高まった。番組を検索する一般的な方法としては、キーワードを入力することにより個人の嗜好に沿った番組検索が可能となっている。しかしながらキーワード入力操作が面倒であり、また、キーワードが個人的な嗜好に偏るため、この方法では、今まで気が付かなかった新しい番組に巡り合うことは稀であった。

そこで、話題の用語や人物を注目キーワードとして自動的に抽出し録画番組検索のためのキーワードとする機能を検討した。

2 注目キーワードと番組選択

番組ストリームと併せ、その番組のＥＰＧメタ情報をデータベースに保存する。ＥＰＧメタ情報の特定期間に着目し、話題性が高いと思われる用語や人物を抽出し、抽出したキーワードを注目キーワードとする。

図 1は注目キーワードを一覧表示した画面（急上昇ワード）の例である。時事ニュースのキーとなる用語、人気が出始めた人物などがピックアップされている。視聴者はリモコンを用いてこのリストから興味あるキーワードを選択する。選択した注目キーワードを検索キーワードとし、一致する番組を録画番組内から検索する。**図 2**は検索した番組を一覧表示した番組選択画面である。視聴者はこのリストの中から好みの番組を選び録画番組を視聴する。

3 システム構成

図 3にシステムの構成図を示す。本システムは、放送信号を番組ストリームとしてストレージに保存し、同時に番組のＥＰＧメタ情報から形態素解析やパターン処理によりキーワードとその意味を取得しデータベースに一時保存する。

注目キーワード抽出処理は、データベースから特定期間のキーワードを取得しキーワード毎に注目度合を求める。システムはこのスコアに基づきキーワードを一覧表示（**図 1**）し、選択したキーワードをもとに番組を検索する。そして検索結果をもとに番組リストの表示（**図 2**）を行い、視聴者が選択した番組を再生する。

4 注目キーワードスコア算出処理の流れ

図 4に示すように抽出期間と対象期間と２つの異なる期間を定め、キーワードと共にその間の出現回数をＥＰＧメタ情報データベースから抽出する。対象期間の出現回数と比較して抽出期間の出現回数が著しく増加しているキーワードが注目キーワードと言える。ところが、ＥＰＧメタ情報から得られるキーワードは、キーワードの出現回数が少なく、差が少ない。このような状況を顧みて、複数のパラメータからなる計算式を用いて、注目キーワードを抽出することとした。ところで、対象期間と抽出期間は視聴者毎に番組録画環境によって調整する。

図 5に注目キーワードスコア算出処理の流れを示す。注目キーワード抽出処理においては、ＥＰＧメタ情報データベースから抽出期間、および、対象期間のキーワード毎の出現回数を取得する。併せてキーワード出現番組時間総計やキーワードの意味を取得する。ただし、ここで番組放送時間が所定以下の番組や再放送等の番組のキーワードは除外する。そしてキーワード毎に5章で述べるスコア計算方法に応じてスコアが求められ、スコア上位のキーワードが注目キーワードとなる。

図 1 注目キーワード表示画面の例

図 2 番組選択表示画面の例

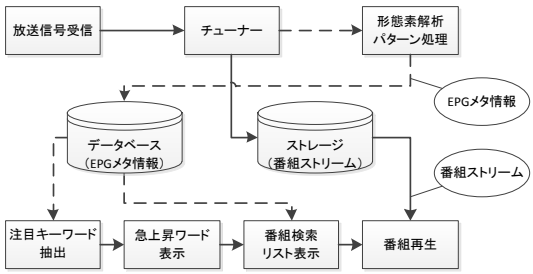

図 3 システム構成図

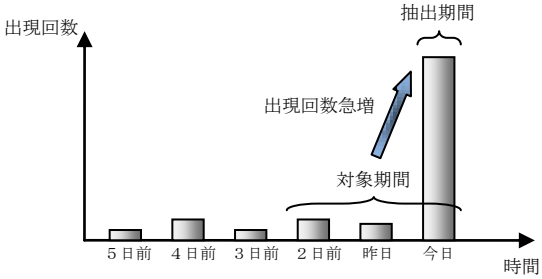

図 4 あるキーワードの時間毎の出現回数分布

5 注目キーワードスコア計算

スコア計算には**図 6**に示す計算式により算出する。以下、各パラメータについて述べる。

5.1 出現回数

抽出期間内のキーワードの出現回数とする。ＥＰＧメタ情報の中に頻出する「日本」「経済」のような一部のキーワードが上位となるが、出現回数が多くても注目キーワードとして不適切なので除外したほうが望ましい。また、大部分のキーワードは狭い範囲内の同一出現回数に偏る傾向があるため、他のパラメータと併用しスコアを計算する。

Evaluation of method to extract Hot-Keyword from EPG metadata
† Michio Yamashita
Platform & Solution Development Center, TOSHIBA Corporation

5.2　出現確率

抽出期間と対象期間のキーワード出現回数の割合により出現確率を求める。この値は、出現回数が多い場合は比較的有意な値となるが、出現回数が少ない場合は高い値が出やすく、変動が大きいため参考程度とするほうが望ましい。この値も特定の値に偏る傾向がある。

5.3　出現番組時間総計

抽出期間のキーワードの出現する番組時間の総計とする。注目キーワードとの相関が出現回数や出現確率に比べ低いと言えるが、出現番組時間総計が長いほど重要なキーワードという考えに基づく。特定の長時間番組に影響されることが多いが、似た出現回数のキーワードの中から注目キーワードを選び出すのに有効となる。

5.4　調整係数

出現回数が少ないキーワードほどスコアを減算する係数とする。この係数は対象期間の出現回数と抽出期間の出現回数から係数値が求まり、結果として、出現回数が少ないキーワードが注目キーワードに選ばれることが抑止される。

6　評価と考察

今回は東京地区地上デジタル放送計６局（ＮＨＫ総合、日テレ、テレビ朝日、ＴＢＳ、テレビ東京、フジテレビ）のすべての番組を３日間連続して録画した環境で評価を行った。対象期間は３日、抽出期間は１日とした。１日あたり、番組放送時間が所定以下の番組と再放送等を除くと約１４０の番組があり、出現キーワード数は約３２００ワードであった。ここから通貨等の数値や、意味区分が判別できない不明語を除くと約１６００ワードとなり、この中から注目キーワードを抽出した。

表 1は、ある日の注目キーワードのスコア上位１０位までのキーワード毎のスコアと各パラメータの順位を示す。出現回数が少ないキーワードは調整係数によりスコアが抑えられるため順位から除外した。キーワード１，２は、出現回数、出現確率が共に高い典型的なパターンの注目キーワードである。出現回数、出現確率の順位が下がるに従いスコア順位も下がる。

表 2は、**表 1**とは別の日の注目キーワードのスコア上位１０位までのキーワード毎のスコアと各パラメータの順位を示す。キーワード１～３は出現回数、出現確率が同一であるが、出現番組時間総計によりスコアに差を出している。キーワード４は出現確率の順位が高いのにキーワード１～３より下位となっている。これは、対象期間の出現回数が少ないことにより、調整係数によりスコアが抑えられるためである。出現確率順位の高いもの(キーワード６)、出現回数順位の高いもの（キーワード８）、両パラメータが程よく高いもの（キーワード５，７）、など特徴的な傾向を持つキーワードが上位にランクインする。

このように複数のパラメータの組み合わせでスコアを計算することにより、ＥＰＧメタ情報の中の種々雑多なキーワードの中からも注目キーワードとして望ましいキーワードを探し出すことが実現できた。また、最適化されたデータベース処理とスコア計算処理を併用することで、軽量な処理負荷のシステムを実現した。

7　おわりに

今回提案する方法にて実装と評価を行い、テレビなどＣＰＵ処理能力が低い機器でも、軽量な処理負荷で有用な注目キーワードが取得できる事を確認した。気になっている物事を扱った番組や人気が出始めた人物が出演する番組を容易に探し出すことはもちろん、世の中のトレンドをいち早くチェックできるようになった。本研究結果をもとに、ＥＰＧメタ情報から注目キーワードを抽出し、番組検索を実現する機能を東芝ＲＥＧＺＡＺ７シリーズに搭載した。

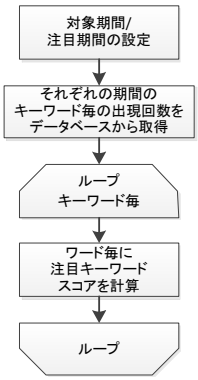

図 5 注目キーワードスコア算出処理の流れ

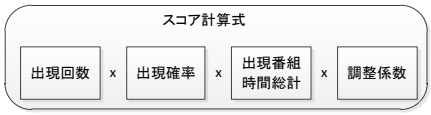

図 6 スコア計算式

ｷｰﾜｰﾄﾞ	スコア順位	各パラメータの順位		
		出現回数	出現確率	出現番組時間総計
ｷｰﾜｰﾄﾞ 1	1	1	9	1
ｷｰﾜｰﾄﾞ 2	2	6	8	2
ｷｰﾜｰﾄﾞ 3	3	8	65	3
ｷｰﾜｰﾄﾞ 4	4	2	66	7
ｷｰﾜｰﾄﾞ 5	5	5	119	11
ｷｰﾜｰﾄﾞ 6	6	15	63	14
ｷｰﾜｰﾄﾞ 7	7	11	66	5
ｷｰﾜｰﾄﾞ 8	8	4	122	9
ｷｰﾜｰﾄﾞ 9	9	15	63	123
ｷｰﾜｰﾄﾞ 10	10	11	66	6

表 1 キーワード毎のスコアと各パラメータの順位(例１)

ｷｰﾜｰﾄﾞ	スコア順位	各パラメータの順位		
		出現回数	出現確率	出現番組時間総計
ｷｰﾜｰﾄﾞ 1	1	6	8	11
ｷｰﾜｰﾄﾞ 2	2	6	8	53
ｷｰﾜｰﾄﾞ 3	3	6	8	126
ｷｰﾜｰﾄﾞ 4	4	6	1	10
ｷｰﾜｰﾄﾞ 5	5	3	47	44
ｷｰﾜｰﾄﾞ 6	6	13	1	12
ｷｰﾜｰﾄﾞ 7	7	13	43	221
ｷｰﾜｰﾄﾞ 8	8	1	90	1
ｷｰﾜｰﾄﾞ 9	9	6	72	401
ｷｰﾜｰﾄﾞ 10	10	13	48	6

表 2 キーワード毎のスコアと各パラメータの順位(例２)

参考文献

[1] 鈴木 優，石谷 康人，坂本 圭，″連鎖検索インタフェース“ササッとサーチ™”″，東芝レビュー，Vol.62，No.12，2007.

[2] 鈴木 優，布目 光生，石谷康人，″インタラクティブなペン操作を可能とする検索意図に基づく連鎖情報検索″，情報処理学会インタラクション 2006 論文集，pp. 101-108，2006.

[3] 菊池 匡晃，岡本 昌之，山崎 智弘，″階層型クラスタリングを用いた時系列テキスト集合からの話題推移抽出″，日本データベース学会論文誌，Vol. 7，No. 1，pp.85-90，2008.