D-036

テレビ番組放送内容に連動した Web 情報選択手法の提案
Information Selection from Web Search Results based on TV Closed Caption Text

中澤 昌美† 帆足 啓一郎† 松本 一則† 滝嶋 康弘†
Masami Nakazawa Keiichiro Hoashi Kazunori Matsumoto Yasuhiro Takishima

1. はじめに

テレビ放送のデジタル放送化やインターネットの普及に伴い，テレビと Web の連携が進められている．テレビ放送を視聴中，番組内容に関連する情報が欲しいことがある．その際，パソコンや携帯などの端末で検索することもできるが，Web 連動テレビを用いるとテレビでの情報検索が可能となり，単にテレビを視聴する場合に比べて，多くの情報を得ることが可能となる．ユーザにとって，検索クエリを自ら考えることが困難であり，入力が手間となっている．そこで，検索クエリを能動的に入力することなく Web 検索が可能な検索手法が考えられている．本研究では，クエリを意識せず，視聴中のテレビ番組に関連する検索を行い，ユーザが欲しいと思う情報を選択して提示する手法について考え，その評価実験を行う．

2. 関連研究

Henzinger ら[1]は，Web から番組の類似ページを検索する手法を提案している．字幕データからキーワードを抽出し，番組の類似ページを検索する．この手法は，テレビと同じ内容のページが欲しい場合に必要だが，番組の詳細情報や他の視点からの情報を得ることができない．また，15 秒ごとに番組を分割しているため，話題単位での情報提示ができないという問題がある．

馬ら[2]は，放送と Web コンテンツの動的統合のための Query-Free 検索機構を提案した．ユーザがアクセスしているコンテンツに基づいて質問を自動生成するため，ユーザが質問を意識することなく検索が可能である．また，テレビ番組の字幕から重要語を抽出し，主題語（subject）と内容語（content）が構成する話題構造により，4 種類（SB：Subject-Broadening，SD：Subject-Deepening，CB：Content-Broadening，CD：Content-Deepening）の質問生成をすることで，単に番組に類似したページを検索するのではなく，番組の内容をより詳しく（Deepening），または別の視点から述べる（Broadening）検索が可能となった．提案手法は，ユーザへの提示対象ページを，4 種類の検索手法の各 1 件目にしている．馬らは，放送中のテレビ番組の補完情報を含む Web ページを得ることを目的としているが，我々の目的は，補完情報からユーザが欲しいと思う情報を選択して提示することである．テレビ番組の話題内容によって異なると考えられる，ユーザが欲しいと思う情報を知ることで，ユーザに提示する Web ページを選択する必要がある．

3. 本研究の目的

本研究の目的は，テレビ番組を視聴する際にユーザが見たいと思う Web 情報を選択する手法を確立することである．本稿では，馬らの Query-Free 検索により関連周辺情報を取得し，4 種類の検索手法の内，どの検索手法を選択するとユーザが欲しい情報が提示できるか調査する．選択の際の一つの指標として，Web ページと字幕との類似度に着目する．Query-Free 検索により取得した関連 Web ページの中には，番組の話題と同じ重要語を含むにも関わらず，全く異なる話題が記載される場合がある．例えば，同じ人物の過去の話題であったり，同じ地域の話題でも，政治であったり事件であったりと，様々な情報が得られる場合がある．視聴したテレビ番組とかけ離れた情報は，ユーザが見たいと思う情報ではないと考えられる．そこで，テレビ番組の字幕と関連情報として検索された Web ページ群をユーザに提示し，適切な Web ページ群を選択する実験を行う．そして，字幕と被験者が選択した Web ページとの類似度を調べる．

†ＫＤＤＩ研究所 KDDI R&D Laboratories

4. アンケート調査

本節では，視聴者が見たいと思う関連 Web ページの調査を行う．実験対象として，2009 年 11 月 5 日から 2009 年 11 月 13 日までの平日 7 日間放送された 45 分間のあるニュース番組を用いる．この番組データから 109 の話題が抽出できた．2 名の被験者に，上記の番組を視聴中と仮定して以下の手順でアンケート調査を行った．

(1) テレビ番組のある話題における字幕情報を抽出する．
(2) 主題度と内容度を算出し，主題語 2 語と内容語 3 語を決定する．
(3) Query-Free 検索手法により 4 種類の質問と，否定の質問を除いた NM 質問を生成し，検索する．
なお，NM 質問は，主題語をタイトルに含み，内容語を本文に含む質問である．ユーザが見たいと思う検索手法の選択において，Query-Free 検索による補完情報検索と NM 検索により得られると考えられる話題と近い内容の検索とを比較するために追加する．
(4) 5 種類の検索結果上位 N 件の URL を取得する．
今回の検証実験では N＝3 とし，各検索手法につき 3 件，合計 15 件の URL を取得した．ただし，検索結果数が 4 件以下の場合はクエリの語数を 1 つずつ減らして検索する．
(5) 被験者に字幕を読んでもらい，見たいと思う Web ページ群を選んでもらう．

以上の手順より，ニュースの話題ごとに視聴者が見たいと思う Web ページが取得できる．

5. 字幕とユーザが好む Web ページの類似度調査

以下に類似度調査を行う手順を示す．

(1) 4.1 節(4)で取得した Web ページから主体となる本文を抽出する．
(2) 字幕と検索結果 URL 本文の特徴データを作成する．
字幕を形態素解析により単語に区切り，tf・idf 値を計算し，重みづけを行った 61844 次元の特徴ベクトルを生成する．また，検索結果ページの特徴データも字幕と同様にして生成する．

表 1．各話題の検索手法選択結果

	SB	SD	CB	CD	NM	計
被験者 1	34	22	16	25	12	109
被験者 2	29	35	27	10	8	109
共通結果	16	10	7	3	0	36

表 2．見たいページと類似度とが一致する話題数

	SB	SD	CB	CD	NM	計
被験者 1	15	9	8	2	3	37
被験者 2	22	15	10	2	3	52

表 3．各検索手法の類似度平均値

	SB	SD	CB	CD	NM	平均
類似度	0.142	0.129	0.122	0.062	0.124	0.117

表 4．各手法の検索クエリの語数

	SB	SD	CB	CD	NM	平均
語数	4.77	4.99	3.43	3.09	2.71	3.80

(3) 字幕と検索結果 URL 本文の類似度を計算する．
生成された特徴量をもとに，字幕と検索結果 URL 間のコサイン相関値を算出する．

(4) 類似度と被験者の選択結果を比較する．
コサイン相関値が最大の Web ページを含む検索手法（SB，SD，CB，CD，NM）と被験者が選択した検索手法を比較する．

6. 調査結果

4 節，5 節で行った調査をもとに，6.1 節ではどの検索手法がユーザに選択されやすいか，6.2 節では類似度とユーザの選択に関連があるのかについてそれぞれ考察する．

6.1 アンケート調査の考察

表 1 に，4 節のアンケート調査結果を示す．2 人の被験者が同じ検索手法を選択した割合が 33.0%と低いことから，人によって見たいページにばらつきがあるといえる．

最も番組との内容が近いと考えて追加した NM 検索の結果得られた情報が被験者に好まれないことから，テレビ番組と似た情報ページを提示しても，視聴者の満足が得られないことが分かる．一旦，周辺の補完情報を得た後，情報を選択して提示するというアプローチは，テレビ番組の関連 Web 情報提示に有効であるといえる．

Subject 系質問（SB，SD）は Content 系質問（CB，CD）より被験者に多く選択された．ユーザは，興味のある Web ページを閲覧するとき，主に見るのはタイトルより本文であると考えられる．ここで，Subject 系質問は，本文に主題語や内容語を含む検索を行うため，見たいと思うページが多くなる傾向にあるいえる．反対に，Content 系質問はタイトルにクエリを含む検索を行う．ユーザが主に見る本文には，クエリとして用いた重要語を直接含まないページが多いために敬遠された可能性がある．

6.2 類似度の考察

表 2 で，被験者が見たいと思うページを選択してもらった結果と，テレビ番組字幕テキストとの類似度が最大の Web ページを含む検索手法を調べた．結果から，被験者 2 に関しては，字幕との類似度が高い Web ページを選択・提示すると，47.7%の確率で見たいと思うページが閲覧できた．しかし，被験者 1 では見たいと思うページが閲覧できる確率は 33.9%と低かった．

表 3 は，各手法の類似度の平均値を示している．NM 検索は，話題に近い情報を検索するため，類似度は高くなると考えていたが，字幕テキストと類似した Web ページが得られる検索にはならなかった．理由としては，NM 検索は，検索結果数が少ないため，クエリの検索語数を減らして検索しており，字幕テキストと似たページの検索が難しくなっているからだと考えられる．一方，CD 検索の類似度は極端に低いが，ユーザが見たいと思う検索として選ばれている．以上より，類似度のみから検索手法を選択することは困難である．

表 4 は，各手法の検索に用いたクエリの検索語数を表す．類似度とクエリの検索語数から，語数が多い検索手法は，類似度が高くなる傾向にある．クエリの検索語数が多いと，その分検索条件が厳しくなるため，話題と内容が似た Web ページが検索される．反対に，クエリの検索語数が減ると，検索条件が緩くなるため，内容が異なる Web ページも検索される． CD 検索の類似度が低い要因としては，クエリの検索語数が少ないことが考えられる．従って，類似度のみを比較するのではなく，クエリの検索語数や，クエリによって得られる Web ページ全体の類似度や検索件数を考慮することにより，ユーザが見たいと思う関連 Web 情報が選択できる可能性がある．

7. まとめ

テレビ番組を視聴する際，ユーザが見たいと思う関連 Web 情報を提示するため，提示すべき Web ページを検討した．本稿では，Query-Free 検索により対象となるテレビ番組の話題の周辺情報を検索し，関連周辺情報を得た後，ユーザが見たいと思う Web ページ検索手法を選択し，提示するというアプローチをとった．ユーザが見たいと思う情報は，テレビ番組の字幕テキストと関連 Web ページとの類似度が高いと仮定し，検証実験を行った．その結果，類似度のみの情報で提示する Web ページを選択することは困難だといえる．今後の課題には，検索結果に含まれる多数の Web ページの中から提示する情報の選択が必要であり，その基準として類似度以外の選択手法を考慮する．

検索手法を選択する際の類似度以外の基準，同じ検索手法の中にも検索結果は多数あるため，Web ページの選定手法を考えるなどが挙げられる．また，一般の人が見たい情報を提示するために， 母数を増やした調査をする必要がある．

参考文献

[1] Henzinger, M., Chang, B.-W., Milch, B. and Brin, S., “Query-Free News Search”, Proc. 12th international World Wide Web Conference (2003)

[2] 馬 強, 田中 克己, “話題構造に基づく放送と Web コンテンツ統合のための検索機構”, 情報処理学会論文誌：データベース, Vol.45, No.SIG 10(TOD 23) (2004).