マルチメディア通信と分散処理 91－21
（1999. 1. 22）

ダイナミックハイパーメディアシステムの構築

～ 感性語によるコンテンツ検索 ～

原田 敦†, 佐藤 克文†, 熊谷 和也†,
鈴木 良宏‡, 上田 謙一‡, 勝本 道哲*, 飯作 俊一*
†(株)松下通信仙台研究所, ‡松下通信工業(株), *郵政省通信総合研究所
E-mail: harada@srd.mci.mei.co.jp

筆者らは，ネットワーク上に分散配置されているマルチメディア情報を，ユーザが容易に検索し，閲覧できる次世代の分散型マルチメディア・プラットフォーム，また，アプリケーションに依存しない汎用プラットフォームとしてのダイナミックハイパーメディアシステムの研究及び開発を進めている．その中で用いられる膨大なコンテンツデータを，感性語を用いて分類する手法を試みたのでここに報告する．

Construction of Dynamic Hyper Media System

- Contents classification by Kansei-word processing -

Atsushi HARADA†, Katsufumi SATO†, Kazuya KUMAGAI†,
Yoshihiro SUZUKI‡, Kenichi UEDA‡,
Michiaki KATSUMOTO*, Shunichi IISAKU*

†Matsushita Communication Sendai R&D Labs. Co., Ltd.
‡Matsushita Communication Industrial Co., Ltd.,
*Communication Resarch Laboratory, MPT
E-mail: harada@srd.mci.mei.co.jp

We proceed with the research for the multimedia platform of the next generation. This platform is composed of distributed multimedia database, and independent of application types. The purpose of this platform is the easy search of multimedia information for anyone. We attempted to classify the multimedia contents which were used in this system, using a Kansei-word processing method. We report a briefing of this experiments.

1. 概要

近年コンピュータ機器やネットワークの高性能化，高速化により，ネットワークを介したマルチメディア・データの扱いが増加している．また，インターネットなどの大規模ネットワークの普及により，情報が広範囲に分散化される傾向があるため，それらの効率的な検索，利用が求められてきている． そこで我々は，ネットワーク上のマルチメディア・データの容易な検索・閲覧を目的とした，分散知識データベースシステムの研究・開発を行なっており，観光案内アプリケーションを搭載したダイナミックハイパーメディアシステムのプロトタイプを構築した．

このプロトタイプで使用している観光案内コンテンツは 1700 個以上に上り，それらを５つの分野に人手を介して分類してある．しかし，コンテンツの分類作業には膨大なコストがかかるため，作業の自動化が求められている．

我々は，ビデオコンテンツの音声データに

含まれている感性語に着目し，それを用いてコンテンツを各分野に自動分類する試みを行なった．

2. 実験方法

DHS に搭載している観光案内アプリケーションでは 10 秒から 5 分程度の動画コンテンツを 1702 個使用している．各々の動画コンテンツには，対応したナレーションの音声コンテンツがある．動画および音声のコンテンツを 5 つの分野に分類した（表 1）．この分類は予め適当と思われる分野を 5 つ設定した上で，各コンテンツを再生したものを実際に人が見て各分野に振り分けた．この作業は女性 2 名により行なった．

表 1 コンテンツの分類とコンテンツ数

分野	主な内容	コンテンツ数
見る	風景,名所	1,031
遊ぶ	レジャー施設	429
食べる	飲食店，食材	135
買う	土産物店	72
泊まる	旅館，ホテル	35
全体	−	1,702

各音声コンテンツデータをテキストデータに変換した．テキストデータに含まれている感性語を抽出し，そのテキストの属する分野毎にその出現頻度を計測した．感性語は形容詞，連体詞とした．感性語を抽出するための形態素解析には茶筅(Ver. 1. 0)を用いた．
次に分野毎の感性語の集合をその分野を特徴づける感性語群とし，感性語群の単語を出現頻度が高い順に並べた集合をその分野の感性語辞書とした．

感性語辞書を用いて全てのコンテンツを再分類した．分類はテキストデータに含まれる単語と分野毎の感性語辞書に含まれる単語とを比較し，そのテキストデータが属する分野を決定する方法をとり，全てのテキストデータについて処理を行なった．

3. 結果

3.1. 感性語の抽出

テキストデータに含まれる感性語の数，種類を分野別に調べた結果を表 2 に示す．ここに，単語数は名詞，動詞，形容詞，連体詞を，感性語に関しては形容詞，連体詞を計測している．

表 2 感性語含有数

分野	単語数	感性語数	感性語種類
見る	32,380	1,679	393
遊ぶ	12,019	672	207
食べる	2,881	247	118
買う	1,511	117	70
泊まる	1,320	89	68
全体	50,111	2,805	489

表 3 感性語含有比率

分野	感性語数比率	感性語種類比率
見る	0.052	0.0121
遊ぶ	0.056	0.0172
食べる	0.086	0.0410
買う	0.077	0.0463
泊まる	0.067	0.0515
全体	0.056	0.0098

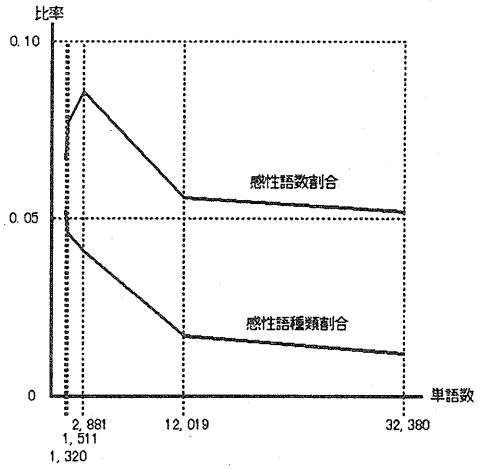

図 1 感性語含有比率

3.2. 感性語辞書

分野別のテキストデータに含まれる感性語を計測し,出現頻度の高い順に並べたものを分野別の感性語辞書とした．各辞書に含まれる感性語を頻度の高い順に 30 個までを表 4に示す．表中の数字は感性語が計測された個数である．
分野別に感性語の出現頻度と感性語の数の比率を図 2に示す．各分野間でほぼ同様のパターンとなっていることが分かる．感性語の数が最も多いサンプルである「見る」分野の感性語出現数のグラフを図 3に，感性語の累積出現頻度のグラフを図 4に示す．

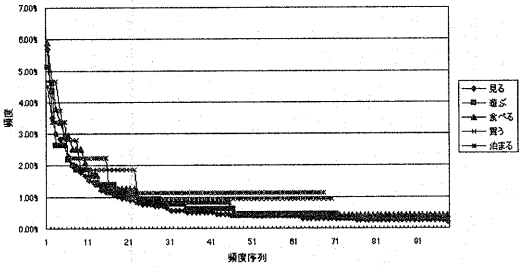

図 2 感性語出現頻度（分野別）

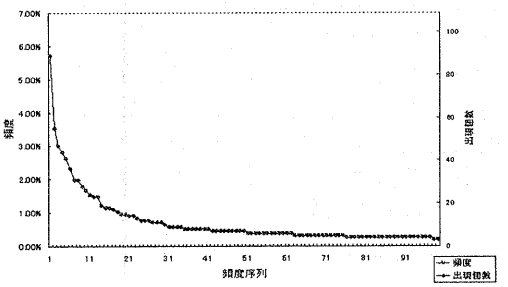

図 3 感性語出現個数（見る）

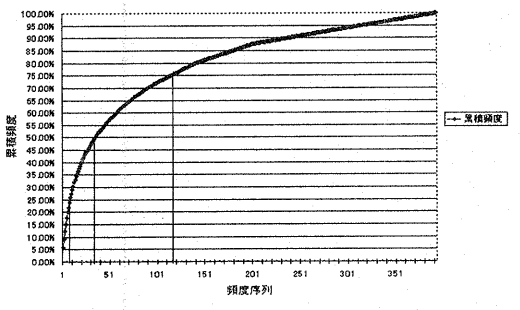

図 4 感性語 累積出現頻度（見る）

3.3. コンテンツの再分類

1 語以上の感性語を含む各テキストデータに含まれる感性語が，各分野毎の辞書に含まれる数を合計する．そのポイントが最もの高い分野にテキストデータを分類した．同ポイントの分野があった場合は分類不可とし，分類失敗と同じ扱いとした．感性語辞書は各分野毎に出現頻度の高い語を n 個含んでいる．n が 5 語から 75 語までの 15 種類の辞書を用いた．結果を図 5に示す．

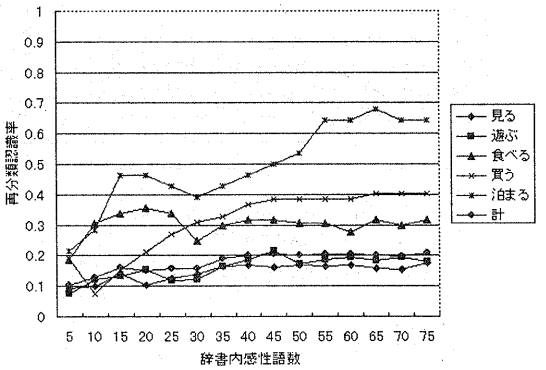

図 5 再分類成功率（重み付けなし）

1 語以上の感性語を含む各テキストデータに含まれる各感性語と，各分野毎の辞書に記載された出現頻度比率との積を合計する．そのポイントが最もの高い分野にテキストデータを分類した．同ポイントの分野があった場合は分類不可とし，分類失敗と同じ扱いとした．感性語辞書は各分野毎に出現頻度の高い語を n 個含んでいる．n が 5 語から 75 語までの 15 種類の辞書を用いた．結果を図 6に示す．

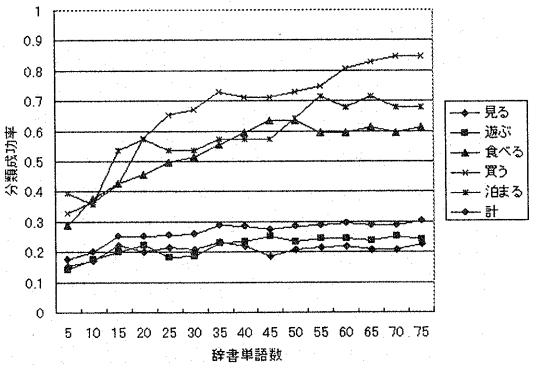

図 6 再分類成功率（重み付けあり）

テキストデータに含まれる感性語の数による再分類結果を図 7に示す．これは図 6で示した感性語に重み付けを行なった再分類結果のうち，全分野の合計の値を用いている．各テキストデータに含まれる感性語の数が 1 個以上のグループから 8 個以上のグループまでの 8 グループをプロットしている．

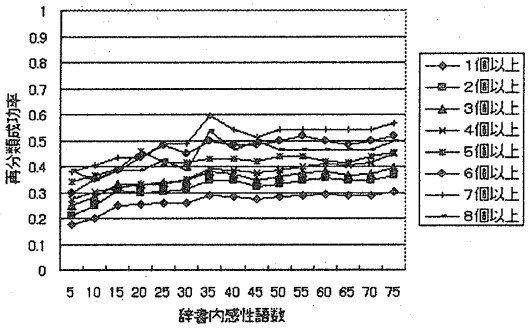

図 7 含有感性語数による再分類

4. 考察

4.1. 感性語の抽出

DHS で使用している 1,702 個のコンテンツの音声をテキスト化したデータを構成する単語（名詞,動詞,形容詞,連体詞）50,111 語から抽出した感性語（形容詞，連体詞)は 2,804 語，489 種になる．各コンテンツの属する 5 つの分野毎に単語数，感性語数,感性語の種類を計測し,その比率をみると，単語数中の感性語数の割合，単語数中の感性語種類の割合とも一定値に漸近しているようである（図 1).

感性語の種類が全単語数に占める割合は，感性語の種類が有限であることを考えると，全単語数の増加に伴いゼロに漸近することになる．本実験のデータでもその様子をうかがうことができるが，全単語数と感性語の種類の比率は 1%程度までのサンプルを用いたということにとどまる．今後，実験を継続または拡張するにあたり，サンプルとなる母集団のプロファイルを表すパラメータとしてこの比率を用いることができるかもしれない．

本実験において感性語の語数が全単語数中で占める割合は 5～9%程度であり，5%程度の割合に漸近しているように見える．ただしこの割合は，観光案内コンテンツのナレーション原稿という性格，もしくはその出筆者の文体等によって決まる値であると思われ，一般のテキストデータに演繹できるかどうかは疑問の余地がある．本実験に用いたテキストデータのプロファイルの一項目として捕らえている．今後，異なる分野のテキストデータにおける同様の比率と比較検討することにより文章プロファイルを特定するパラメータとして利用できると思われる．

4.2. 感性語辞書

分野毎のテキストファイルに含まれる感性語を抽出し,頻度順に並べたものを感性語辞書とした．各辞書の感性語が元のテキストデータの中で出現する頻度の比率は分野間で差違がない（図 2)．また，グラフ（図 3, 図 4）より，出現頻度の高い少数の感性語でその分野を特徴づけられることが分かる．サンプル数の多い「見る」の分野において見てみると，出現頻度上位 9 個の感性語で出現比率の 25%をカバーし，35 個の感性語で 50%の出現比率をカバーする． 35 個の感性語は「見る」分野に出現した感性語の種類数 393 の 9%にすぎない． また表 4より，各分野辞書に含まれる感性語は，多少重複しているものの出現頻度の序列まで含めて考えると各分野を十分に特徴づけていると考えられる．

4.3. コンテンツの再分類

DHS において使用している 1,702 個の音声コンテンツをテキストデータ化したもののうち,感性語を 1 語も含まないものが 518 個あった（表 5)．これは観光案内コンテンツの性格上，オーディオのデータが音楽のみで文章を含まないものや,名詞を羅列した説明を行なうコンテンツが多いためである．感性語を用いたコンテンツの再分類実験においては，1 語以上の感性語を含むテキストデータ 1,184 個を用いて行なった．

4.3.1. 重み付けをしない感性語辞書

感性語を含む各テキストデータから感性語を抽出し,感性語辞書に含まれる感性語の数

を分野毎の辞書に対して計測し,最もポイントの高い分野にそのコンテンツを分類した.感性語辞書に含まれる感性語数を 5 語から 75 語まで 5 語刻みで実験をした.(図 5)「泊まる」,「買う」,「食べる」という分野の再分類成功率が他と比べて高くなっているが,これらの分野はコンテンツ数が少ないため,コンテンツから作成した感性語辞書を利用し,そのコンテンツを再分類したために成功率が上がっていると考えられる.コンテンツ数の多い「見る」分野においては成功率が 20%弱程度であり,辞書内の感性語数の差による再分類成功率は 35 語以上であれば優位な差はない.全体的に見ても,辞書の単語数を 35 個以上とすれば,再分類成功率は 20%程度であった.

4.3.2. 重み付けをした感性語辞書

各テキストデータから感性語を抽出し,感性語辞書に含まれる感性語の数と分野毎のその感性語の出現頻度の積を,分野毎に計測し,最もポイントの高い分野にそのコンテンツを分類した.感性語辞書に含まれる感性語数を 5 語から 75 語まで 5 語刻みで実験をした(図 6).
重み付けなしの実験に比べ,全体的に成功率は向上している.
重み付けなしでの実験で比較的高い再分類率であった,「泊まる」,「買う」,「食べる」という分野の成功率は,35 語以上の辞書を用いた場合 6〜80%程度に高まっている.しかし,この数値自体は,重み付けなしの実験で述べたように,これらの分野のコンテンツ数が少なさに起因することは否めない.

4.3.3. 含有感性語数による再分類結果

重み付けをした感性語辞書を用いて全分野のテキストデータを再分類する際に,テキストデータに含まれる感性語の数をパラメータにしたグラフが図 7である.全体としては感性語を多く含むテキストデータほど再分類成功率が高い傾向がある.今回用いたデータでは,感性語を 7 個以上含むテキストデータの数が少ないため(表 5),4〜6 個の感性語を含むテキストデータに関して 4~50%の再分類成功率を得られ,1 個以上の感性語を含むテキストデータでも 30%程度の分類成功率が得られた感触である.

5. まとめ

今回用いたテキストデータは平均すると 1 データあたり 28.8 語の単語(全テキストデータ)と 2.4 語の感性語(感性語を含む 1,184 個のテキストデータ)を含んでいる短いデータであったにも関わらず 4 割程度の分類成功率を得られた.また,作成した分野毎の感性語辞書毎の感性語出現頻度のプロファイルに分野間の差が見られない(図 2)こと,また,比較的少数の感性語でその分野が特徴づけられる可能性があること(図 3,図 4)を考えると,本手法は,より多くの単語を含むテキストデータを分類する際には有効な手段となり得ることが期待される.
今後,単語数の多いデータでの実験,辞書間の重複単語の扱いを考慮した判定方法の改良,辞書作成時とは異なるサンプルを用いた分類検証を通して本手法の確立を目指すとともに,別の手法と組合せての分類成功率向上をはかる予定である.

参考文献

[1] 阿比留,他:“映像を中心とした分散知識データベースシステムの構築”,情報処理学会第 54 回全国大会講演論文集(3),3Q-3(1997)
[2] 佐藤,関,音喜多,鈴木,上田,飯作:“映像を中心とした分散知識データベースシステムの構築(2)観光案内システムへの適用”,情報処理学会第 54 回全国大会講演論文集(3),3Q-4(1997)
[3] 佐藤,熊谷,音喜多,阿比留,鈴木,上田,勝本,飯作:“映像を中心とした分散知識データベースの構築” 第 84 回マルチメディア通信と分散処理研究会,Sep. 1997
[4] 原田,熊谷,佐藤,鈴木,上田,勝本,飯作:“分散知識データベースの高機能化” 情報処理学会第 55 回全国大会講演論文集(3),6G-04(1997)
[5] 原田,熊谷,佐藤,鈴木,上田,勝本,飯作:“ダイナミックハイパーメディアシステムの構築 -プレゼンテーション制御言語の設計-” 第 85 回マルチメディア通信と分散処理研究会,Nov. 1997
[6] 原田,熊谷,佐藤,鈴木,上田,勝本,飯作:“ダイナミックハイパーメディアシステムの構築” 情報処理学会第 56 回全国大会講演論文集(3) 4Z-07-9,1998

[7] 佐藤,原田,熊谷,鈴木,上田,勝本,飯作："DHSにおける個人情報管理エージェント" 情報処理学会第 57 回全国大会講演論文集(3),6G-04(1998)

[8] 原田,熊谷,佐藤,鈴木,上田,勝本,飯作："ダイナミックハイパーメディアシステムの実装" マルチメディア通信と分散処理ワークショップ ,IPSJ Symposium Series Vol.98, No.14,pp7-12.(1998)

表 4 感性語辞書（上位 30 語）

見る		遊ぶ		食べる		買う		泊まる	
美しい	5.71E-02	美しい	5.14E-02	新鮮だ	5.91E-02	いかがだ	4.67E-02	贅沢だ	4.49E-02
有名だ	3.53E-02	自然だ	4.36E-02	美味しい	4.64E-02	独特だ	4.67E-02	白い	3.37E-02
古い	3.01E-02	良い	2.65E-02	ある	3.80E-02	ある	4.67E-02	豊かだ	3.37E-02
自然だ	2.82E-02	様々だ	2.65E-02	有名だ	3.38E-02	華麗だ	3.74E-02	ふんだんだ	3.37E-02
高い	2.63E-02	楽しい	2.65E-02	ない	2.95E-02	様々だ	2.80E-02	素晴らしい	3.37E-02
ある	2.31E-02	有名だ	2.18E-02	甘い	2.95E-02	自然だ	2.80E-02	便利だ	2.25E-02
豊かだ	1.99E-02	高い	2.02E-02	良い	2.53E-02	良い	2.80E-02	最適だ	2.25E-02
様々だ	1.99E-02	いっぱいだ	1.87E-02	いかがだ	2.53E-02	有名だ	2.80E-02	美味しい	2.25E-02
多い	1.80E-02	新しい	1.87E-02	独特だ	2.53E-02	広い	1.87E-02	ビッグだ	2.25E-02
良い	1.67E-02	珍しい	1.87E-02	生だ	2.11E-02	繊細だ	1.87E-02	明るい	2.25E-02
貴重だ	1.54E-02	いかがだ	1.87E-02	最高だ	1.69E-02	我が	1.87E-02	ある	2.25E-02
およそ	1.48E-02	沢山だ	1.71E-02	美味だ	1.69E-02	優しい	1.87E-02	様々だ	2.25E-02
大きな	1.48E-02	およそ	1.40E-02	格別だ	1.69E-02	愛らしい	1.87E-02	いっぱいだ	2.25E-02
見事だ	1.22E-02	多い	1.40E-02	大きな	1.27E-02	素朴だ	1.87E-02	心地よい	2.25E-02
広い	1.15E-02	美味しい	1.40E-02	おしゃれだ	1.27E-02	オリジナルだ	1.87E-02	自然だ	2.25E-02
我が	1.15E-02	健康だ	1.40E-02	少ない	1.27E-02	優美だ	1.87E-02	古い	1.12E-02
静かだ	1.09E-02	ある	1.40E-02	およそ	1.27E-02	懐かしい	1.87E-02	静かだ	1.12E-02
白い	1.03E-02	華やかだ	1.25E-02	素朴だ	1.27E-02	少ない	1.87E-02	高い	1.12E-02
新しい	9.62E-03	白い	1.09E-02	充分だ	1.27E-02	ふわだ	1.87E-02	別だ	1.12E-02
雄大だ	9.62E-03	青い	1.09E-02	豪快だ	1.27E-02	微妙だ	1.87E-02	至便だ	1.12E-02
青い	8.98E-03	ない	1.09E-02	大きい	1.27E-02	見事だ	1.87E-02	小さな	1.12E-02
鮮やかだ	8.98E-03	大きな	1.09E-02	沢山だ	1.27E-02	盛んだ	1.87E-02	独特だ	1.12E-02
豪華だ	8.34E-03	多彩だ	9.35E-03	自然だ	8.44E-03	きらびやかだ	9.35E-03	重だ	1.12E-02
広大だ	7.70E-03	気軽だ	9.35E-03	珍しい	8.44E-03	暖かい	9.35E-03	充分だ	1.12E-02
素朴だ	7.70E-03	自由だ	9.35E-03	多い	8.44E-03	数多い	9.35E-03	数多い	1.12E-02
平安だ	7.70E-03	メインだ	9.35E-03	盛んだ	8.44E-03	大胆だ	9.35E-03	良い	1.12E-02
数多い	7.06E-03	豪快だ	9.35E-03	とがだ	8.44E-03	楽しい	9.35E-03	近代的だ	1.12E-02
安全だ	7.06E-03	熱い	9.35E-03	爽やかだ	8.44E-03	豊かだ	9.35E-03	抜群だ	1.12E-02
新鮮だ	7.06E-03	長い	7.79E-03	秘密だ	8.44E-03	精密だ	9.35E-03	まずい	1.12E-02

表 5 感性語含有数によるテキストデータの分布

分野	感性語含有数																			計
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
見る	326	295	172	102	64	34	15	5	8	2	2	1	1	1	0	2	1	0	0	0	1,031
遊ぶ	130	125	81	34	27	24	4	2	1	1	0	0	0	0	0	0	0	0	0	0	429
食べる	35	41	30	8	9	6	2	2	0	0	0	1	0	0	0	0	0	0	0	1	135
買う	20	18	21	5	3	3	0	1	1	0	0	0	0	0	0	0	0	0	0	0	72
泊まる	7	9	7	4	1	1	2	1	2	0	1	0	0	0	0	0	0	0	0	0	35
計	518	488	311	153	104	68	23	11	12	3	3	2	1	1	0	2	1	0	0	1	1,702