2V-1

感性データ解析による カテゴリ検索システムの概念体系

四ッ谷雅輝　　石本展啓　　三浦孝夫

法政大学工学部電気電子工学科

1　はじめに

現在の WWW には，各ユーザにとって有益な情報が多々存在するが，不必要な情報もそれ以上に存在する．この膨張した WWW から欲しい情報だけを得るために，情報検索システムは重要な役割を担っている．その現在のオンライン情報の検索システム（サーチエンジン）では，ユーザが求める検索結果を得るのは困難であるというのが現状である．

原因の一つとして，現在の Web ページはキーワードという一つの要素で絞り出せるほど，小規模ではないからである．また，未知の事柄を検索する上で，その時の思いつきで与えたキーワードが適切である保証はどこにもない．そこで，本研究では，感性という新しい要素を取り入れることにより，現在の情報検索における問題点の解決を試みる．この問題点の解決により，我々が得られるユーザ支援の具体的内容は，二つに大別できる．

一つ目は，より洗練された検索結果の獲得である．一般的に，Web ページからユーザが欲しい情報を検索する時，頼りにするのは片言のキーワードのみである．ユーザはその情報については未知であるが，サーチエンジンにキーワードを入力する時には，ユーザの頭の中には欲しい Web ページのおよその概観は浮かんでいることが多いだろう．曖昧ではあるがその時想像した概観は，膨大なデータからの絞込みを試みる上で，重要な情報になり得る．キーワードによって絞り込み，感性情報によりもう一段階絞り込むことができたら，より高品質な検索結果を得ることが可能になるだろう．

二つ目は，新たな角度からの検索法の発見である．すなわち，新しいキーワードの発見である．感性がもたらす情報は，主観的で曖昧ではあるが，その本質を説明するかもしれない．感性情報は本人が言葉として具体的に表すことができなかった心底にある要求への対応の貴重な情報になるだろう．この対応は，本人さえも意識していなかった新たなキーワードを発見し，違う角度からの検索という手段を与える．これにより，ユーザは選択の幅が与えられ，意図する検索結果を得られる可能性は向上するだろう．

2　検索システムの構成

本研究で提案するシステムでは，事前に WWW 上から無作為に抽出したサンプル Web ページを因子分析及びクラスタ分析を用いて，Web ページを特徴化できる感性ワードを抽出する．その各感性ワードに対してスコアを計算し，キーワードとともにデータベースに格納する．ユーザは感性ワードの評価尺度の選択入力と未知の情報に対する片言のキーワードを入力すると，本システムはデータベースに格納された各スコアとユーザの感性との近傍探索を実現し，検索結果として複数個のキーワードを返す．これを既存のサーチエンジンに自動的に渡し，URL を獲得する．この工程を順を追って説明する．

2.1　感性ワードの取得

事前に無作為に抽出したサンプル Web ページ (約 100 件) に対して，30 個のパラメータを与え，WWW の Web ページの特徴を捉えることを試みる．パラメータは Web ページをまず，文学的作品という観点から捉え，使われている文字や品詞の種類に対して 12 個制定した．続いて，Web ページは HTML で記述されていることに注目し，使われている機能（タグ）や関連技術 (Java Script など) に対して 18 個制定した．この 30 種類のパラメータを実測し，データ行列を生成する．

本研究では，各 Web ページを評価する重要な部分を担う，感性ワードを抽出する方法として因子分析とクラスタ分析を用いる．これらの分析手法を用いるねらいは，変量間の相関を考慮し，多変量を少ない変量に凝縮できると考えたからである．

先述のデータ行列を主因子法による因子分析により，因子得点行列を得る．そして，得られた因子得点行列を Ward 法によるクラスタ分析により，各パラメータをクラスタリングする．得られたその結果を図 1 に示す．図 1 より，感性ワードとして，「叙述的な」「大々的な」「簡潔な」「派手な」の 4 つを採用する．

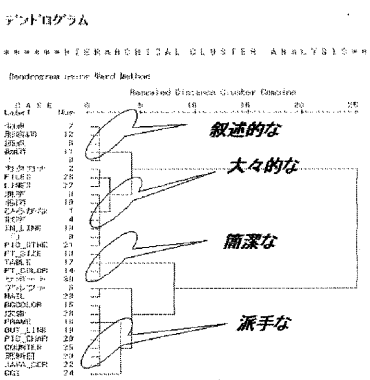

図 1: 各因子得点のクラスタ分析の結果

Conceptual Design of Category Search Systemusing Kansei Data Analysis
Masaki Yotutani, Nobuhiro Isimoto, Takao Miura
Hosei University, Dept.of Elec. and Elec. Eng.
Kajino-cho 3-7-2, Koganei, Tokyo, JAPAN

2.2 スコアの付与

前節より，各 Web ページにおいて 30 個のパラメータで表していた特徴を新しい 4 つのパラメータを導入することによって，特徴を損なわないように表現する．言い換えればこの 4 つのパラメータは 30 個のパラメータを代表しており，各 Web ページの特徴を凝縮していると言える．具体的には，4 つ感性ワードは「叙述的な」「大々的な」「簡潔な」「派手な」の順に「4」「11」「4」「11」個の各パラメータから成り立っている．これに基づき，感性ワードのスコアを次式で与える．スコアを導入するねらいは，各 Web ページの特徴を定量化することである．スコアはキーワードと共に各 Web ページごとにデータベースに格納され，次節の近傍探索で利用される．

$$W_{score} = \frac{1}{n}\sum_{k=1}^{n} P_k = Z_i \quad (1)$$

$$P_k = \frac{x - \bar{x}}{s} \quad (2)$$

P_k ：構成するパラメータ $(1 \leq k \leq n, n = 4, or\ 11)$
Z_i ：Z 得点 $(1 \leq i \leq 4)$　s ：パラメータの標準偏差
x ：パラメータの値　$\bar{x}$ ：パラメータの平均

2.3 近傍探索

入力された片言のキーワードをカテゴリとみなし，データベース内のリレーション名とマッチさせる．これにより，検索対象となるリレーションを決定する．次に，前節により得られたスコアとユーザから選択入力された感性ワードの評価尺度を数値化し，近傍探索を実施する．これは，ユーザの感性とサンプル Web ページの特徴との差異（距離）を計算し，検索半径に応じて値を返すという検索方法である．この時用いたスコアから入力評価尺度までの距離は次式で与える．

$$d = \sqrt{(Z_1 - U_1)^2 + (Z_2 - U_2)^2 + \cdots + (Z_n - U_n)^2} < r \quad (3)$$

U_i ：評価尺度の変換数値　r ：検索半径 $(0.2 \leq r \leq 2)$

3 実験と結果

サッカーの Web ページが欲しい時に，感性情報が検索結果に与える影響を実験する．各 Web ページのスコアとテスト入力した感性との距離を表 1 に示す．テスト入力の感性のコンセプトは，A は「こじんまりとしていて，すっきりとした」で評価尺度の入力は，$U_1 = 0.3, U_2 = -0.9, U_3 = 0.9, U_4 = -0.3$ であり，B は，「大々的で複雑な」で評価尺度の入力は，$U_1 = 0.3, U_2 = 0.9, U_3 = -0.9, U_4 = -0.3$ である．この時の $r = 1.4$ での検索結果を表 2，3 に示す．*judgment* とは，”サッカー”とユーザに提案するキーワードと関連性が妥当であるかどうかを判断したものである．

Category = ”サッカー”

	score				*d*	
ID	Z_1	Z_2	Z_3	Z_4	*A*	*B*
1	-0.036	-0.026	0.357	0.305	1.240	1.708
2	-0.267	-0.426	-0.319	-0.244	1.427	1.556
3	-0.314	-0.277	0.066	-0.089	1.227	1.655
4	0.858	0.350	0.101	0.868	1.969	1.726
5	-0.463	-0.566	-0.278	-0.256	1.443	1.766
6	1.356	1.293	2.126	1.803	3.442	3.853
7	-0.189	-0.280	-0.383	-0.072	1.524	1.397
8	0.081	-0.209	-0.345	-0.290	1.441	1.259
9	-0.298	-0.144	-0.357	0.192	1.659	1.409
10	-0.418	-0.605	-0.380	-0.351	1.498	1.747
11	-0.424	-0.277	-0.376	0.332	1.715	1.607
12	-0.030	-0.164	-0.376	0.153	1.576	1.312

表 1 ： 各 Web ページのスコアとテスト入力した感性との距離

ID	*keyword*	*judgment*
1	フットサル	△
	試合	△
	広島県	×
3	サッカー用品	○
	ワールドサッカー	○
	山口県	×

表 2 ： A の検索結果

ID	*keyword*	*judgment*
7	審判	○
	フットサル	△
	審判用具	△
8	キリンカップ	○
	ブラジル	△
	日本代表	△
12	草サッカー	○
	対戦相手募集	△
	メンバー募集	△

表 3 ： B の検索結果

A の検索結果では，いずれの Web ページも階層も浅く，小規模であった．一方 B では，階層構造が複雑で，A よりもユーザとのインタラクトを意識しており，BBS などの機能もあり，規模が大きかった．

表 1 より，A, B の各 ID について，Z_2 では，目立った傾向を得ることはできなかったが，Z_3 に着目すると，$ID = 1, 3$ では，値が正で大きく，$ID = 7, 8, 12$ では，値が負で小さくなっている．よって，本システムで導入したスコアは概ね各 Web ページの特徴を捉えることができた．また，表 2 より，サッカーに関連性があるキーワードをユーザにいくつか提案することができた．

問題点としては，各スコアの散らばりの幅が狭いことが挙げられる．原因として，目測でしか得られないパラメータのため，実測できる Web ページの種類に限界あり，サンプルとして取り上げた時点で，Web ページのある程度の画一化が成されたと考えられる．同じ理由からデータベースに格納してあるサンプルデータが小規模であるため，因子空間が存在しないところもあり，検索結果を得られない感性の入力も存在してしまっている．

次に，各スコアの散らばりに偏りが生じていることも挙げられる．これにより，ユーザがヒット数と検索半径の対応を把握するのは困難になっている．また，各 Web ページの大まかな特徴しか捕らえることができず，Web ページの個性を色濃く反映させるには至らなかったと言える．原因として，感性ワードの生成時に，誤って分類されているパラメータの存在が考えられる．例えば，「簡潔な」という感性ワードを構成しているパラメータの一つに”サポート言語の数”がある．これは，Web ページが日本語以外にも他の言語で読めるかを示すパラメータである．このパラメータを採用した目的は，この数値が高いページほど，公的要素または学術的要素が高いということを示す一つの目安になると考えたからであるが，全く予期していない「簡潔な」という感性ワードに分類がなされた．その理由として，当てあまる Web ページが少なくパラメータにならなかったことが考えられる．本研究では，開発者の主観の混入の懸念から，各分析の結果をダイレクトに感性ワードに結び付けたが，このようなノイズを手作業でも，統一されたコンセプトに基づき除去すれば，洗練された感性ワードを生成することができるだろう．また，有意性の議論や信頼度係数の導入により，さらに洗練された感性ワードの取得を期待することもできる．

4 結び

本論文では，感性情報により，情報検索システムの強化を図り，ユーザ支援を実現するシステムを提案して，情報検索における感性情報の有効性を検証した．本システムでは，従来の検索方法では考慮することができなかった，ユーザの感性を検索に反映することができた．従って，検索結果として得られた Web ページは個人差はあるが，ユーザの嗜好にある程度応えることができる．また，関連するキーワードをいくつか提示することにより，新しい切り口での検索の提案を可能にし，検索において必要とされるユーザの専門知識を補助することができた．

今後の展望として，対象データの拡大が考えられる．方法の一つとして，WWW 上の任意の Web ページのパラメータを収集できるロボットの構築を実現し，ロボットによって得られた大規模なデータベースを定期的に更新することができたら，ユーザの感性をより直接的に，情報検索に結びつけることができる感性サーチエンジンの構築に繋がるだろう．

参考文献

[1] 安本 美典 著: 因子分析法，培風館 (1981)

[2] 柳井晴夫 共著: 因子分析，朝倉書店 (1990)

[3] 中森 義輝 著: 感性データ解析，森北出版 (2000)

[4] 北川博之 著: データベースシステム，昭晃堂 (1996)