

感性データ解析によるカテゴリ検索システムの設計と実装

山崎 悦弘 川島 俊二 四ッ谷 雅輝 三浦 孝夫

法政大学工学部電気電子工学科

概要

現在の web ページ検索において、ユーザはあらかじめ何らかのキーワードを持っていないければ、サーチエンジンは使い物にならない。つまり、キーワードがなければ検索できないという点に大きな問題がある。本論文では、ユーザの感性より感性キーワード生み出す感性 web システムの設計と実装について述べる。この感性 web システムによりサーチエンジンを大きく強化することができる。というのは、このシステムが既存のサーチエンジンに組み込みが可能であるからである。

1 前書き

現在、サーチエンジンそのものを作ることに目が向けられている。しかし、本研究では既存のサーチエンジンをうまく利用するという考えで、サーチエンジンの拡張を提案する。つまり、サーチエンジンにキーワードを投入する前処理を考える。既存のサーチエンジンにおいてはキーワードを元に検索をしている。もし、そのキーワードが思い浮かばなかったらどうなるだろう。例えば、曲名の知らない歌を検索したい場合などである。その問題を解決するために、我々は感性 web システムを設計する。感性 web システムとは、キーワードが浮かばないユーザにユーザ自身の感性からキーワード（ヒント）を与えるシステムである。以降、混乱を避けるため感性 web システムにより得られるキーワードを「感性キーワード」とする。もちろんキーワードをあらかじめ持っているユーザでも、より幅の広い検索を行うためにもこのシステムを利用できる。以上のことにより、この感性 web システムは既存のサーチエンジンをより強化でき、有効に使うことができる。

2 動機と解法

ある曲を検索したい場合、その曲の歌詞のイメージや曲のイメージはなんとなくだがわかる。しかし、肝心の曲名が浮かばないとき、ユーザはどのように検索するだろうか。解法として、歌詞の一部をキーワードとして検索する。歌詞がその曲の核の部分となっていればヒットするかもしれないが、いつもそれが核となっているという可能性は極めて低い。次の解法として、ユーザのもつイメージからジャンル（ポップロック演歌）を想像し、キーワードとして検索する。一見よさそうであるが、ヒットがかなり多くなってしまふ事が予想される。前

挙げた 2 つの解法だと結局のところ問題を解決できない。最終的な解法として、以下を提案する。何らかの手段でヒット件数をかなり絞れるキーワードを用いて検索する。例えば、キーワードとしてアーティスト名まで絞れば、ヒット件数から見てもかなり有効的である。ここでの何らかの手段に感性 web システムを用いる。

3 システム構成

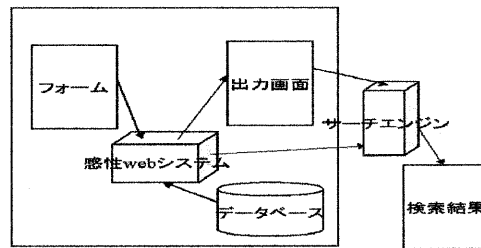


図 1: 感性 web システムの流れ

システムの流れは図 1 のようになる。キーワード、ユーザの感性、検索範囲をフォームに入力し、感性 web システムにアクセスする。それを感性 web システムにおいてデータベースにアクセスし、検索範囲にある感性キーワードを抽出する。抽出された感性キーワードをサーチエンジンに投入し、その結果を画面に表示する。

3.1 フォーム

ユーザの感性を入力するフォームは HTML により作成し、ユーザの感性を取り込むためにラジオボタンを使用する。ラジオボタンは感性ワードの尺度を 7 段階として設定する。これにより、ユーザの感性を数値化する。ラジオボタンによりユーザの感性尺度を感性 web システムに送る。また、より良い感性キーワードの抽出のために検索範囲をユーザに指定してもらう。

3.2 感性 web システム

感性 web システムは CGI に比べてサーバへの負担の少なく、データベースへのアクセスが容易である PHP を使用する。SQL により、データベースにアクセスする。近傍探索により、フォームから入力されたユーザの感性尺度と因子得点を比較して、ユーザの感性に近い感性キーワードを取得する。PHP ファ

イルにはラジオボタンにより、設定したユーザの感性和因子得点との距離 d を計算するために $\$SQL=update$ テーブル名 set $d=(Z1-\$U1)*(Z1-\$U1)+(Z2-\$U2)*(Z2-\$U2)+\dots+(Zn-\$Un)*(\$Un-\$Un)$ として直接 SQL を記述する。 $\$SQL=select$ keyword from テーブル名 where $d<r*r$ を記述することにより、検索範囲半径 r 内の感性キーワードを抽出する。また、PHP ファイルに $\langle meta\dots \rangle$ を記述することにより、ユーザに感性キーワードを表示させずに、抽出した感性キーワードを直接サーチエンジンに投入する。このパターンは検索速度を重視する。また、 $pg_exec()$ を用いてキーワードのタプルを返してもらうことにより、抽出した感性キーワードを表示し、サーチエンジンに投入するパターンも用意する。これにより、抽出された感性キーワードを確認することができる。使用するデータベースには、感性キーワードと因子得点のみを格納してある。

3.3 近傍探索

フォームにより、ユーザが指定した感性的なイメージを n 次元空間における座標として捉える。ユーザの指定した座標を $(U1,U2\dots Un)$ として、ユーザの指定した座標を中心とした半径 r の範囲内の任意の点 $(Z1,Z2\dots Zn)$ を収集する。 r はフォームにおいてユーザが入力している。ユーザの指定した座標から半径 r の範囲内の点の距離を d として

$$d = \sqrt{(Z_1 - U_1)^2 + (Z_2 - U_2)^2 + \dots + (Z_n - U_n)^2} < r \quad (1)$$

を感性 web システム内で計算し、 d の昇順に表示する。

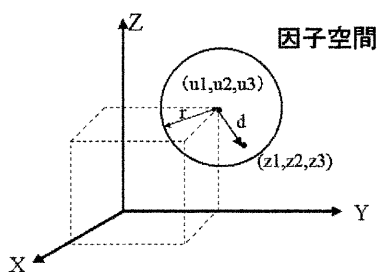


図 2: 近傍探索のモデル図

4 実験

本研究では、フォーム、感性 web システム、データベースは同一計算機上で動作させる。負荷を分散させるために複数の計算機上でも動作は可能である。一連の作業は UNIX 上で行う。データベースに Postgres を使用する。まず、ブラウザ上のフォームにおいて感性ワードに対する 7 段階のラジオボタン 4 個からユーザの感性を入力し、同時に検索範囲も入力する。実験では図 3 に示すように、キーワードとして「歌」を与え、ユーザの感性をすべて中間値近辺を指定し、検索範囲は $r=0.5$ を指定する。最後に感性 web システムで抽出された感性キーワードを出力画面として表示する場合と表示しない場合のどちらか一方を選択する。それを感性 web システムにおいてデータベースにアクセスし、検索範囲にある感性キーワードを抽出する。抽出された感性キーワードを出力画面に表示させる場合を図 4 に示す。図 4 で示すように「オリジナル曲、音楽、MIDI、スタッフコンサート、スケジュール」の 5 個が感性キーワードとして抽出される。抽出した感性キーワードをサーチエンジンに投入し、図 5 で示すように 4 件ヒットを表

示する。図 4 が示すように $r=0.5$ で検索した場合、感性 web システム内部で行われる近傍探索により、具体的なキーワードが得られる。感性 web システムはサーチエンジンの前処理を行うことにより、サーチエンジンとの共存を実現する。それにより、曖昧なイメージを持つユーザにキーワードを与えることを可能にし、サーチエンジンの強化を実現する。

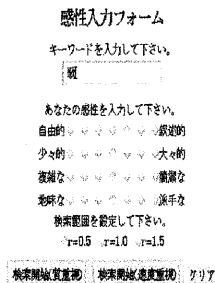


図 3: 入力画面



図 4: 感性キーワード出力結果



図 5: 検索結果

5 評価

$r=1.0$ で検索した場合、感性キーワードが 39 個抽出した。しかし、ほとんどの感性キーワードはユーザにとって、不要なキーワードであった。これは $r=1.0$ での近傍探索では検索範囲が $r=0.5$ の場合と比べて大きく、因子の絞込みが困難になり、因子得点の大きいものまで収集しているからである。一般に、検索範囲が大きくなれば、主要因子への依存度が急激に弱まり、選んだ因子以外の効果を配慮しなくなる。このため、一定の近傍以外は、本稿で提案するアプローチは効果的ではない。

6 結び

本研究ではユーザの感性を 1 点しか取らなかったが、2 点を取って論理積の内部で検索をかけると、1 点のときよりも精度の高い検索結果が得られるかもしれないと推測している。同様に、ラジオボタンの感性ワードの尺度の範囲を細かく設定、感性ワード (n 次元) の数を増やすとより精度の高い結果が得られるかもしれないと推測される。後者の問題点としては、ラジオボタンの数が何個でユーザの感性ワード選択の限界に達するかということである。

参考文献

- [1] 北川博之: データベースシステム, 昭晃堂 (1996)
- [2] 中森義輝: 感性データ解析, 森北出版 (2000)