1B2-NFC3-2

インテリジェントアイコンによる Web 検索結果閲覧支援

Browsing support of Web search results with Intelligent Icon

西海 俊秀[*1]
Toshihide Saikai

高間 康史[*1]
Yasufumi Takama

[*1] 首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

This paper proposes an intelligent icon for supporting the browsing of Web search results. Ordinary search engine result pages (SERPs) employ a linear list of ranked URLs. In such SERPs, various types of Web pages are sorted based on common criterion, which makes it difficult to find desired pages when those are ranked lower in the list. In order to solve such a problem, snippets and thumbnails of web pages are used as a clue to judge the relevance of pages by browsing SERPs only. Aiming to provide more various aspects of retrieved pages, the proposed intelligent icon summarizes and visualizes several features extracted from Web pages. This paper describes the concept of the intelligent icon with some preliminary experimental results.

1. はじめに

サーチエンジンで一般に広く利用されている検索結果提示方法は，ランク付けされた URL の線形リストであるため，タイプの異なる多様なページが共通の指標に基づき一列にソートされてしまう．そのため，求めるタイプのページがランキング上位になかった場合，多種多様なページが混在したリストの中からそれらを見つけ出すことは困難な作業となる．実際のページにアクセスして確認していたのでは検索に時間がかかるため，個々のページにアクセスするか否かの手がかりとなる情報を検索結果ページ(search engine result page, SERP)に提示する必要がある．現在多く用いられている手法がスニペットであるが，ページ内の局所的な情報に過ぎないため，ページの持つ多様な特徴を表示することはできない．Google ではページのサムネイル画像も確認できるようになっているが，具体的なコンテンツに関する情報を提示するものではない．

本稿では，ページの持つ多様な特徴を要約したアイコン(インテリジェントアイコン[Keogh 06])を生成し，ページ選択の手がかりとする手法を提案する．各ページから 16 種類の情報を抽出して特徴量を計算し，4×4 のセルに配置する．各セルは特徴量に応じて配色し，類似した特徴を持つページのアイコンは類似するようにすることで，ページにアクセスするか否かの判断を支援する．提案手法とスニペットを組み合わせた検索結果ページを作成し，スニペットのみを用いた検索結果ページとの比較を通じて，提案手法の有効性や特性について考察する．

2. 関連研究

2.1 検索結果ページの改善に関する研究

サーチエンジンの検索結果には多くのページが含まれているため，全てのページにアクセスしていたのでは時間がかかってしまう．そこで，ページにアクセスするか否かの手がかり情報を提供することが重要となりつつあり，現在はスニペットが一般に用いられている．スニペットとは，検索されたページ中からクエリを含むテキストを抜粋したものであり，一種の要約とみなせる．スニペットは，ページにアクセスすることなく大まかな内容を把握することができるが，クエリ周辺のテキスト情報以外は表示されない．

また，モバイルメディア向けの手がかり情報提示手法も研究されている[斉藤 01]．デスクトップ型の PC と携帯電話を使い分けることを想定し，詳細な情報を閲覧する場合は大画面の PC を利用するが，閲覧するか否かの判断は携帯電話を用いて行う．この論文では，株価の情報を攪拌凝縮法を用いて可視化している．攪拌凝縮とは，各カテゴリに対応する色相で色づけした点を，ランダムかつ稠密に矩形領域に配置する操作である．この手法を用いてポートフォリオの情報を可視化することで，全体的な色合いに基づき，一瞥して安定か不安定かを判断することができる．

2.2 インテリジェントアイコン

インテリジェントアイコンは，1 次元データ列を図1のように 2 次元セルにマッピングする手法である[Keogh 06]．各セルの色は，各属性の値に対応している．属性値は全て正規化されており，図 1 では，0～1 に正規化されている．データの内容が類似するファイルのアイコンは視覚的に類似するため，関連するファイルを容易に見つけることができる．

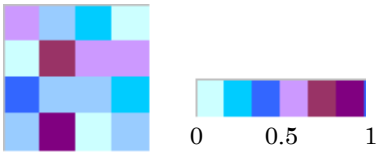


図 1. Intelligent Icon の例

3. 検索結果閲覧支援のためのインテリジェントアイコン

本稿では，Web ページから，そのページにアクセスするかどうかの手がかりとなるようなインテリジェントアイコンを生成することで，SERP の閲覧を支援する事を試みる．インテリジェントアイコンのサイズは 4x4 とし，各ページからその特徴を現す 16 種類の情報を抽出し，ページごとにアイコンを生成する．スニペットがページの局所的な情報のみを提示するのに対し，提案手法で

連絡先：高間康史，首都大学東京大学院システムデザイン研究科，東京都日野市旭が丘 6-6，ytakama@sd.tmu.ac.jp

はページ全体の情報を集約して提示可能なアイコンを生成する事を試みる. 提案手法は, Web ページからの特徴抽出処理とアイコン生成処理に分類される. 以下では各処理について説明する.

3.1 特徴抽出

本論文では, 図2に示す 16 種類の情報を提示する. これらはテキストから抽出される情報とそれ以外に大別できる.

テキストからは 10 種類の特徴を抽出する. 名詞の中でも特徴的な単語である固有表現として, 地名, 人名, 組織名, 日付表現, 時間表現, 金額表現, 割合表現の 7 種類についてページ中での出現回数を求める. 図2において白色の特徴が固有表現に対応する. 左上にある黄色の特徴はテキストから抽出されるその他の情報である. 「クエリ」はクエリの出現回数を指す. クエリと共起している全名詞の中で一番出現回数の高い名詞の占める割合を求め, これを「共起」特徴とする. この割合が高いページは, ある一つの事柄に重点をおいた内容だと考えられる.

この他, 検索された各ページの文書ベクトルを求め, 全ページの重心ベクトルとのコサイン類似度を特徴「コサイン」として用いる. 類似度が低ければ, 検索結果中の他のページとは異なった内容を持つ特徴的なページと言える.

テキスト以外から図2中にオレンジで示されている 6 種類の情報を抽出する. リンクと画像は, それぞれページ内に含まれる個数を特徴とする. 文字数はタグを含まないページの文字数, サイズはソースファイルのサイズ, タグは HTML タグ数を表す. 文字数の多いページはテキスト情報が, ファイルサイズが大きい, あるいはタグの多いページは, テキスト以外の情報が多いページである可能性が高いと考えられる. URL は, ページ URL の「/」の数を抽出する. 「/」の数が多いほどトップページから離れているページだとわかる.

コサイン	クエリ	リンク	画像
共起	人名	組織名	日付表現
文字数	地名	時間表現	割合表現
タグ	ファイルサイズ	金額表現	URL

図2. 手がかりとして利用するページの特徴

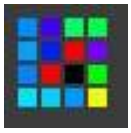
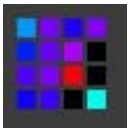
3.2 アイコン化

取り出した 16 種類の特徴を 4×4 のセルに配置し, 特徴量に応じて色を割り当てる. 配置は図2に示す通りである. 関連があると考えられる特徴を隣接させ, 関連の薄い特徴を離れるように配置した. 色は 8 段階とし, 黒, 青, 黄, 赤の順に, 暗い色から鮮やかな色へと変化する. 特徴量が小さいほど黒に近づき, 大きいほど赤に近づく. 実際のアイコンを図3に示す. (a)のページはファイルサイズや文字数が多く情報量が多いため, 全体的に赤が多くみられる. (b)のページは, 全国の海に関するショップ等の情報を掲載しているため, 地名や組織名が多く見られ, 対応するセルが赤くなっている. (c)のページは, 全体的に情報量が少ないため, 青っぽい色調になっている.

4. 評価実験

スニペットのみの検索結果と, スニペットと提案手法を組み合わせた検索結果を比較する実験を行った. 検索エンジンは, Yahoo!API を利用した. 「山」と「海」をクエリとして検索した結果について, 一方をスニペットのみ, もう一方は提案手法を組み合わせた検索結果を用い, 6 人の被験者に Web ページを探すタスクをしてもらった. 検索結果の上位 25 件を提示し, 1 件目のページと似ているページを残りの 24 ページから探す時間を計測した. また, 実験後に, アイコンの有無により検索の仕方にどのような違いがあったかについて質問した.

提案手法を用いた検索結果の場合は, 被験者全員, 色調の似ているアイコンからいくつかの候補を挙げた後, それらのスニペットから最終的な結果を出す傾向がみられた. 図3の例の場合は, (a), (b)のアイコンが似ていると判断された. これにより, 候補をいくつかに絞れるため, 5 人の被験者は提案手法の方が早く検索することができた. 残る 1 人は, アイコンが類似するページは存在しないと判断し, 全ページのスニペットを見直したため, 提案手法を用いた場合の方が遅くなる結果となった. 検索時間が最も短縮された被験者では, 17 分 5 秒から 10 分 10 秒と, 6 分 55 秒の短縮となった. 他の被験者は 1～3 分程度の検索時間で, 30 秒前後の短縮となった. 逆に遅くなってしまった被験者は, 1 分 22 秒から 2 分 12 秒と, 50 秒遅くなる結果となった. 以上より, ほとんどの被験者において提案手法は有効であり, またスニペットと組み合わせることで有効に機能することがわかる.

(a)  (b)  (c) 

(a) 海 – Wikipedia
http://ja.wikipedia.org/wiki/%E6%B5%B7
(b) 海 Explorer（海に関する総合リンク集）
http://www.next-explorer.com/sea/
(c) 沖縄美ら海水族館-沖縄の神秘をありのまま-
http://oki-churaumi.jp/

図 3. 実際のアイコン例

5. おわりに

本研究では, ページにアクセスするか否かの手がかりを提示する事を目的として, ページから抽出した 16 種類の特徴をインテリジェントアイコンとして提示する可視化手法を提案した. 被験者実験でも示した通り, スニペットとは異なる特性を有するため, 両者を組み合わせることで検索結果ページの改善に貢献すると考える.

参考文献

[斉藤 01] 斉藤康彦:モバイルマルチメディア社会の到来 携帯電話の小画面における情報の視覚化に関する一考察, 情報管理, 44(12), 846-854, 2001.

[Keogh 06] Eamonn Keogh, Li Wei, Xiaopeng Xi, Stefano Lonardi, Jin Shieh, Scott Sirowy, ”Intelligent Icon: Integrating Lite-Weight Date Mining and Visualization into GUI Operating Systems,” ICDM’06, pp912-916, 2006.