Web 上の HTML 文書を用いた 意外性のある情報の獲得支援

野口 大輔 (710056)

北陸先端科学技術大学院大学 情報科学研究科

2009 年 2 月 5 日

キーワード: Web 検索, データマイニング, 上位下位関係, GUI, 類似語.

本稿では，Web 上の HTML 文書を解析した結果を元にした，意外性のある情報をユーザが得るための検索キーワードの想起を支援する方法を調査する．また，その具体的な方法として,「鳥式」ユーザインターフェースの改善と，トピックと関連語間の情報を元にした意外性評定の異なる二つの方法で，ユーザの「鳥式」で用いられる関連語における「意外性のある情報」の獲得支援を提案する．

近年の情報化社会における急激な技術発展に伴い，我々一般のユーザがインターネットに触れる機会も日常的なものとなった．その中で，検索エンジンを用いて情報を得るということも，また有り触れた光景である．あるキーワードに関する問題回避や，あるいは行動に関する未知のアイディア，Tips について情報を求めようとする場合，ユーザが「意外」と思うようなキーワードを入力しなければならない場合がある．例えば，健康補助成分「DHA」が挙げられる．これは，過剰に摂取すると，血液の凝固作用を阻害し出血しやすいといったトラブルを引き起こす要因となるものである．あらかじめ検索を行う際に「DHA」に加えて「出血」という意外なキーワードを入力すると，検索結果上位に問題の事実が見つかる．このような意外なキーワードというものは，その関連がユーザの「知識の範囲外」であるがゆえに，何らかのシステムによって「気付かせる」ことが必要ということである．

このようなキーワードの想起を支援するため，我々の研究室では「鳥式」と呼ばれる検索ディレクトリの開発を行ってきた．これは，ユーザが最初に入力したキーワードに対して，関連語を意外な物まで含めて「トラブル」「方法」「ツール」という意味的カテゴリに分類して提示し，検索に利用できるようにするものである．

「鳥式」の既知の問題点として，あるトピックについて検索を行った場合，検索結果として得られる関連語がトラブル表現だけで数十，数百に上る場合がある．また,「鳥式」において検索された関連語は，表示順が Web 文書上でのトピックと関連語との共起頻度を元にしたスコアでソートされているのみであり，ユーザは全ての関連語に目を通さないと,「意外性のある情報」に辿り着けない可能性がある．

そこで，本研究ではユーザインターフェースの改善と，トピックと関連語間の情報を元にした意外性評定の異なる二つの方法で，ユーザの「鳥式」で用いられる関連語における「意外性のある情報」の獲得支援を提案する．このようなトピックと関連語に基づいた「情報の意外性」について研究した例は他にない．

ユーザインターフェースの改善については，

1. FLASH を用いたユーザインターフェースの二次元グラフ化
2. 共起頻度に着目したノード配置
3. 類似度に着目した擬似的ノードクラスタリング
4. 上位下位関係を応用した検索拡張を行う．

１. について,「鳥式」ユーザインターフェースを，FLASH をベースにしたものに変更する．２. について，意味的カテゴリと関連語との距離 r は，トピックと関連語が Web 文書中に現れる際の共起頻度を元に正規化したスコアを用いる．３. について，配置角度を類似度に基づいて決定する．ここでの類似度は，任意の単語と単語の間の類似度に比例してスコアを与えられたものである．４. について，上位下位関係を用いた上位概念による関連語の「継承」を実装する．

実験では，システムの妥当性に対して 70%の肯定的な意見が得られた．

トピックと関連語間の情報を元にした意外性評定では，

1. トピックと関連語の頻度情報に着目したランク付け手法
2. 相互情報量に着目したランク付け手法
3. 単語間類似度に着目した類義語によるランク付け手法
4. 関連語内の順位に着目したランク付け手法を行う．

１. について，世間一般で知名度が高い情報ほど，それに関して記述がなされている文書数が多く，あまり知られていない情報（つまり，意外な情報）ほど文書数が少ないということに着目する．２. について，相互情報量により，トピックと関連語の結び付きの強さを評価する．相互情報量の値により順位づけを行うことによって，出現頻度に関係なく，結び付きの強い言い回しを得る．３. について,「意外性のある情報」とは，あるトピックが属するであろう集団には，そのトピックに付随する関連語が，その集団において珍しい存在だということを表している．ここで，あるトピックが属するであろう集団は，トピックと類似度の高い語群と言い換えることが可能であると考える．４. について，あるトピックと関連語のペアについて，同一の関連語を持つトピック群の中で，相互情報量で降順ソートした場合に，上位に来るものは意外な情報が多いと考える.「そもそも未知の関連語」に属する組が上位に多くなる．

トピック	トラブル表現	解説
綿	農薬被害	綿への「枯葉剤」の付着
綿	チリダニ	綿ぼこり中のハウスダストの原因
歴史劇	濫造	中韓の高句麗論争
商標権	希釈化	「味の素」などへの商標の一般化
商標権	消滅	更新忘れによる権利の消滅
脳膿瘍	感染性疾患	細菌が血液によって脳に到達し，脳の中に膿が溜まる
日本国歴代内閣	犯罪	講和条約「日本国憲法」について
陰陽師	暗殺	暗殺者としての陰陽師
工場廃水	スラッジ	廃水処理殿物（専門用語）
カーボン	付着する	船舶に見られる「黒い汚れ」
カーボン	削りカス	車で，アルミリムの削りカスがカーボンリムを傷つける
樹林	乾燥化	サンショウウオ減少の要因
ひまわり	灰色かび病	専門用語
ひまわり	黒斑病	専門用語
ブータン	児童売買	インドの国境近くで児童売買
蜂	糞公害	養蜂場周辺での汚染

表 1: 獲得できた意外な情報の例と解説

実験では，提案した手法を適用することにより，Average Precision，R Precision において 23%の精度が最高 34%程度まで向上することを示した．また，各手法により表 1 のような既知，未知の関連語を含む「意外な情報」を得ることができた．