

ユーザのニーズに適合した検索語支援システム

関根 裕昭[†] 東 基衛[‡]

早稲田大学大学院 理工学研究科 経営システム工学専攻[†]

1 研究背景と目的

現在, WWW (World Wide Web) 上には膨大な数の Web ページが存在しており, その中から効率的に自分の欲しい情報を見つけることは容易ではない。既存の検索エンジンはキーワード検索が主流であるが, 探している分野について不慣れなユーザにとって検索対象をキーワード化することは容易ではない[1]。そのため, 検索語支援の期待が高まっている。そこで, 本研究ではユーザの過去の興味や嗜好からユーザニーズの推定を行い, 検索語を生成するシステムの研究を行う。ユーザニーズに適合した検索語を新たに提示することで, 検索語の想起に寄与することを目的とした。

2 現状分析と問題点

現在, 検索語支援のサービスは実用化されている。代表的なものに, Google サジェストがある。検索全体での人気度を考慮し, 多数のユーザに頻繁に検索された語を優先的に提示する仕組みである。また, 検索語支援に関する研究もさかんに行われている。これらの問題点を以下にまとめた。

- 1) 推薦される検索語が万人向けである。
- 2) 推薦される検索語が限定的である。
- 3) ノイズが多い。

3 研究アプローチ

2 章で述べた問題点を解決するシステムを実現するため, 以下の3つのアプローチを採る。

1) ユーザプロフィールの作成

従来の検索語支援システム (以下, 現在のシステム) は, 推薦される検索語が万人向けのため個人のニーズに適合した検索語を支援することができない。本研究では, 各ユーザの過去の興味を表したプロフィールをサーバ上に作成することで解決した。提案システムでは個人のニーズを汲み取って検索語を支援することが可能になった。

2) 検索語の生成に検索結果をフィードバック

現在のシステムは, 他人が入力した検索語しか推薦されない。本研究では, Google から取得した検索結果の Web ページを利用し, 検索語を生成することで解決した。他人が入力した検索語以外にも幅広く検索語を生成することが可能になった。

3) 文書のフィルタリング

現在のシステムは, 検索語を生成する際に上位文書をフィルタリングせず利用しているため, ノイズが入る。本研究では, 閲覧時間からユーザの興味の有無を判断し, 検索語生成に利用する文書のフィルタリングを行うことで解決した。

4 提案システム

4.1 提案システム概要

提案システムの概要図を図 1 に示す。

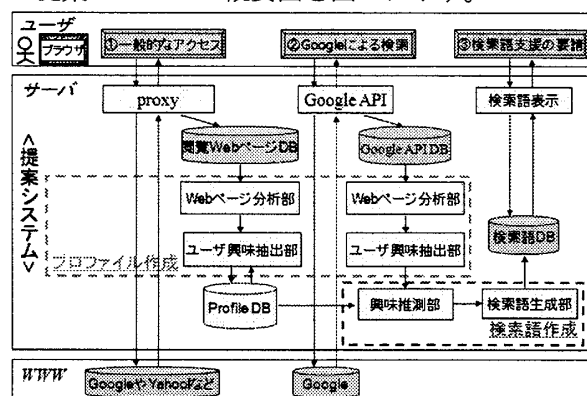


図 1 提案システム概要図

提案システムは, 図 1 に示すように, 4つのコンポーネントとデータベース, ユーザインターフェイスから構成される。ユーザの作業は以下の3つである。①一般的なアクセス: 検索語の推薦を必要としない場合の WWW へのアクセスのことである。②Google による検索: ブラウジング中に検索語の推薦を必要とする場合の WWW へのアクセスのことである。③検索語支援の要請である。

次に, ユーザの作業に対応して, システムがどう動作するかを述べる。以下はシステムの動作である。①一般的なアクセス: ユーザにプロキシサーバ経由で Web へアクセスしてもらい, 閲覧履歴・閲覧時間を取得する。②Google による検索: ユーザに Google API を使って, Web へアクセスしてもらい, ユーザの現在の興味を取得する。③検索語支援の要請: 上記の①及び②を経過した段階で, 検索語支援が可能になる。②で得られたユーザのその時のニーズと①で得られるユーザの今までの興味や嗜好を考慮し, 検索語を生成する。

4.2 プロファイルの作成

<Web ページ分析部>

ユーザの閲覧履歴及び閲覧時間をプロキシサーバから取得し, Web ページの分析を行う。

i 閲覧時間を計測し AND 検索用と NOT 検索用

An Internet Search Support System Guided by Interpreting User's Intention

[†]Hiroaki SEKINE, [‡]Motoei AZUMA

Dept. of IMSE, Graduate School of Sci. and Eng., Waseda Univ.

に閲覧ページを分類する

(ア) 閲覧時間 $t \geq t_0$ のとき: AND 検索用

d_p : 興味のあったページの単語集合

(イ) 閲覧時間 $t \leq t_0$ のとき: NOT 検索用

d_M : 興味のなかったページの単語集合

t_0 は, ユーザの閲覧行動を分析し決定する。

ii 形態素解析を行い, 単語をそれぞれ抽出する

(ア) AND 検索用

$$\vec{d}_p = (w_1(D_p), w_2(D_p) \dots w_N(D_p)) \quad (1)$$

(イ) NOT 検索用

$$\vec{d}_M = (w_1(D_M), w_2(D_M) \dots w_N(D_M)) \quad (2)$$

ここで, $w_i(D)$ は文書 D に含まれる単語である。

<ユーザ興味抽出部>

Web ページ分析部から, 分析結果をもらい, ユーザの興味を抽出する。以下 $X = P$ or M とする。

iii AND 検索用, NOT 検索用それぞれ別に単語の出現回数をカウントする

$$d_x = w_1(P\{10\}, M\{5\}), w_2(P\{3\}, M\{0\}), \dots \quad (3)$$

$$\dots, w_N(P\{k\}, M\{l\})$$

ここで, $w_i(P\{k\}, M\{l\})$ は, i 番目の単語 w が, \vec{d}_p で k 回, \vec{d}_M で l 回出現したことを表す。

iv 単語の偏り具合 β を利用し, AND 検索用と NOT 検索用の両方に含まれる語を削除する

$$\beta = \frac{|k-l|}{\sqrt{k^2+l^2}} \quad (4)$$

v ユーザプロフィール U の設定

$$U_x = \sum_i^N w_i(P\{k\}, M\{l\}) \quad (5)$$

ここで, $X = P$ のとき正のプロファイル, $X = M$ のとき負のプロファイルと定義する。

vi ユーザプロフィール U の更新

$$U_x = U'_x + \sum_{i=0}^N w_i(P\{k\}, M\{l\}) \quad (6)$$

ここで, U'_x は過去のユーザプロフィールである。

4.3 検索語の生成

推薦される検索語の範囲を図2に示す。本研究では, 図2の中で I, II, III に含まれる単語を検索語として支援する。

I 集合 N : 今までに目にしたことのない単語を支援 (集合 G とユーザプロフィールとの差分)

II AND 検索用として検索語を支援 (正のプロファイルと現在の興味の共通部分)

III NOT 検索用として検索語を支援 (負のプロファイルと現在の興味の共通部分)

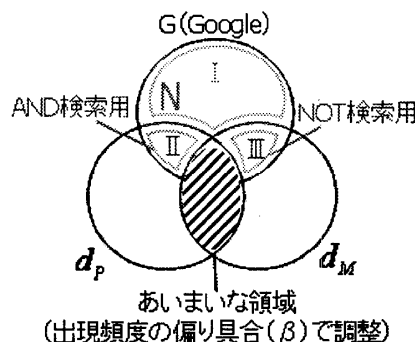


図2 推薦される検索語の範囲

集合 G : 現在の検索結果から抽出した単語集合
<検索語生成部>

AND 検索用, NOT 検索用の検索語を生成する。

1. ユーザの初期検索語 S を Google で検索
2. 検索結果の上位 T 件を抽出
3. 形態素解析を行い単語を抽出 (集合 G とする)
4. 興味推測部により, 単語に得点をつけ, 得点の高い順にユーザへ提示

<興味推測部>

ユーザの今までの興味・嗜好を考慮し, 現在のニーズを推測する。

1. ユーザプロフィール U_x に含まれる単語と集合 G に含まれる単語の共通部分を抽出
2. 抽出した単語に重み (α) をつけ, 得点とする
単語の得点は以下の式で決定される。

$$w_i = (\alpha U_x \{u\}, G\{g\}) = \alpha u + g \quad (7)$$

ここで $U_x \{u\}$ 及び $G\{g\}$ はそれぞれユーザプロフィール, 集合 G での単語 W の出現回数である。

5 評価実験と考察

本研究の有効性を検証するために, プロトタイプを利用して評価実験を行った。定量的評価として, 推薦検索語の適合率を測定したところ, 高い適合率が得られた。さらに, 定性的評価として実施したアンケートにおいても過半数以上から肯定的な回答が得られた。

6 結論と今後の課題

本研究では, 個人の閲覧履歴からプロフィールを作成し, 推薦検索語のパーソナライズ化を行った。さらに, Web ページから検索語を生成することで, 他人が入力した検索語以外にも対応できるようになった。今後の課題として, 適合率の向上及び利用シーンの拡張が挙げられる。

参考文献

- [1] 吉岡真治ら: “検索語の網羅性に注目した汎化概念により検索語選択支援を行う情報検索システムの研究”, 人工知能学会論文誌, 20 巻 4 号 A(2005), pp.270-280