

検索セットの共有による検索支援システムの提案

草場 裕雄†

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

Web 上には膨大な情報が存在し、増え続けている。一般の人々がその中から情報を探するとき、検索エンジンを用いる。しかし、Google などに代表される検索エンジンは必ずしも個人に最適化されたサービスではない。そのため近年では個人に適応した検索支援を行う研究が盛んになっている。

一般にユーザが Web 上の情報を検索する作業の問題点として検索効率やユーザ個人の熟練度合いや検索対象に対する知識に左右されることが挙げられる。

本研究では、ユーザが検索時に入力した情報や、検索過程において興味を持ったサイトに基づいて検索ニーズを把握する。そして同じニーズどうしで検索結果を組織内で共有させ、効率よく目的の情報源に導くことが目的である。

2. 研究目的と用語の定義

検索のコンテキスト

検索クエリのコンテキストは研究者によって定義が異なっている。検索のコンテキストを利用する意味は、ユーザの検索ニーズをより詳細に把握することにある。本研究では、ユーザが検索する際に閲覧した Web ページの集合を検索のコンテキストと定義する。

利用状況の想定と研究目的

検索エンジンを用いた Web 上の情報収集ではユーザが独立して検索を行うため情報を共有することが困難であり、効率的な情報収集ができない。そこで、本研究では協調フィルタリングの仕組みを用いて知識の共有を行う。協調フィルタリングは人間の情報収集行動から興味・関心・意図を把握し、共通の問題意識を持った者に情報提供することで情報収集活動を支援する技術であり、有効な手法である。

3. 従来研究とシステム要件

SERF[1]と AntWorld[2]はユーザが過去に入力した検索クエリを検索ニーズとし、過去に似たクエリから良い評価の Web コンテンツを推薦する。

両研究はクエリ（自然文）同士を完全一致あるいは、ベクトルを用いて類似度計算しているため、検索のコンテキストを考慮に入れて類似度を計算していない。よってクエリという少ない情報量を基に完全一致による類似度計算は精度に問題が生じる。従来のシステムでは、ユーザの興味を把握する段階で以下のような問題点が存在した。協調フィルタリングの問題点として①推薦する Web コンテンツの内容を考慮に入れていない。また、SERF などの研究では②検索クエリのコンテキストを考慮に入れていないことが挙げられる。以上の問題点を考慮し、以下3項をシステムの要件とする。

1. 検索する際のコンテキストを考慮に入れる。
2. 他者の情報を用いて効率よく Web ページを推薦する。（目的より）
3. コンテンツの内容とユーザの評価、両方を考慮に入れて推薦を行う。

4. 研究アプローチ

3のシステム要件に対して以下のアプローチを行う。

1. 検索時に検索クエリとカテゴリを入力。
2. ユーザと似た検索ニーズから得られた好評価の Web ページを共有。
3. 検索過程において閲覧したコンテンツをベクトルで表現し、さらにユーザ評価も考慮。

5. 提案システム

アルゴリズム

本研究のシステムではユーザが検索をする行動からユーザの嗜好を推測する。検索結果から Web ページをクリック、あるいは Web ページを評価するたびにベクトル（Track ベクトル）が更新される。推測された嗜好とカテゴリをもとに、他のユーザの検索セットと類似度を計算し、Web ページを推薦する。本研究では検索ニーズの比較を行う。検索ニーズはユーザが検索時に入力したクエリ、カテゴリ、Web ページの内容、タイトルも含むサマリ、Web ページに対する評価（検索セット）によって表現される。検索セットの類似度を式1に示す。類似度が高い検索セットの中で、良いと評価された Web ページ、クリックされた Web ページを他のユーザに推薦する。

$$simSearchSet = simTrack + simInput \quad (1)$$

Track ベクトルはユーザの検索行動（検索結果のページから閲覧した Web ページ）を表現し、*Input* ベクトルはユーザが入力したクエリ、カテゴリを表現する。

Track はユーザの行動に基づいて 4 つの文章集合に分類される。その後に 4 種類の *Track* ベクトルを比較し、*simTrack* が計算される。

Input はクエリ、とカテゴリからベクトルが生成され、ベクトルを比較して *simInput* が計算される。

Web ページとサマリの分類

本研究では Web ページの内容とそのサマリをベクトルで表現する。それらはユーザが検索過程においてとった行動に基づいて D_1 から D_4 に分けられる。

D_1 : ユーザに**良い**と評価された Web ページの集合

D_2 : 検索結果からユーザに**選ばれた**サマリ（タイトルも含む）の集合

D_3 : ユーザに**悪い**と評価された Web ページの集合

D_4 : 検索結果からユーザに**意図的に選ばれなかった**サマリの集合

D_2 と D_4 の分類ではユーザは検索結果を上から順に閲覧して選択するという仮定に基づいて行われる。

Track ベクトルの計算

Track ベクトル $\vec{d}$ は Web ページの文書のベクトル、あるいは Web ページのサマリのベクトルである。 $\vec{d}$ は上で示したルールに基づいて 4 種類に分類され、それぞれの文書集合 D_1 から D_4 の平均ベクトルが算出される。 $\vec{d}$ は単語の頻度を文書全体に出現する総単語数を割った TF 値によって得る。集合の平均ベクトルどうしで類似度を計算し *Track* の類似度が算出される。類似度はコサイン相関値を用いて計算する。ユーザの検索行動に基づくベクトルの類似度を式 4 に示す。

$$simTrack(Track_1, Track_2) = \sum_{i=1}^4 sim(aveDoc_{1i}, aveDoc_{2i}) / 4 \quad (4)$$

$aveDoc_i$: 文書集合 i に含まれる文書ベクトルの平均ベクトル

$simTrack(Track_1, Track_2)$: *Track* ベクトル 1 と *Track* 2 のコサイン相関値

Input ベクトルの計算

Input ベクトルはカテゴリの単語と入力キーワードから生成されるベクトルであり、各文書との TF 値から求められる。検索セットを計算する際、*Input* ベクトル

の平均ベクトルが計算される。*Input* どうしの類似度を算出するにはコサイン相関値を用いる。

プロトタイプ



図 1 検索結果

このプロトタイプでは検索セットの記録とコンテンツ推薦を行う。初期の検索画面でユーザはキーワードとカテゴリを入力し、検索を行う。検索結果の画面で上部にシステムの推薦ページが表示され下部に Google の検索結果が表示される。検索結果から選択されたページは随時 DB に保存される。Web ページを表示する際には上部にキーワードを変更して検索を行うためのテキストボックス、検索ボタン、そしてコンテンツに対する適合、不適合の評価が行えるボタンを表示している。評価した後は DB に評価が記録され、再び検索結果が表示される。

6. おわりに

本研究では検索ニーズを検索クエリ、カテゴリ、閲覧した Web ページとそれらのページに対する評価から推測した。そして、類似した検索ニーズの間で良いと評価されたページを共有する手法を提案した。

今後の課題として、アルゴリズムの改良、ユーザに対する負荷の軽減が考えられる。

参考文献

- [1]. Seikyung Jung, Kevin Harris, Janet Webster, Jonathan L. Herlocker. "SERF: integrating human recommendations with search" Proceedings of the 13th Conference on Information and Knowledge Management November 8-13, (2004)
- [2]. Kantor, P.B., Boros, E., Melamed, B., Menkov, V., Shapira, B. and Neu, D.L. Antworld: Capturing Human Intelligence in the Net. Communications of the ACM, 43, 8, (2000)