ユーザの検索意図を考慮したクラスタリング検索システム

小部山 知伸† 東 基衛†
早稲田大学大学院 創造理工学研究科 経営システム工学専攻†

1. はじめに

膨大な情報の中から有用な情報を探すために、検索エンジンを利用することが一般的である。しかし、適切なキーワードを入力することが困難であるということや不要なWebコンテンツ(以後、コンテンツ)が含まれてしまうという問題がある。そこで本研究では、ユーザの検索意図に適したコンテンツを容易に収集するクラスタリング検索システムを提案する。

2. 現状分析と問題点

クラスタリング検索システムの問題点として以下の2つをあげた。

1)コンテンツの特徴が十分に反映されず不要なコンテンツが含まれやすい

従来、コンテンツ全体から特徴的な語を抽出し、コンテンツベクトルを表現している。しかし、一つのクラスタに不要なコンテンツが多数含まれてしまう。また、頻出する語の影響を受けやすく、頻出しない特徴的な語が抽出されにくい。

2)クラスタリング結果にユーザのフィードバック情報が十分に反映されない

検索結果へのフィードバックを行うために、適合・不適合判定したコンテンツまたはクラスタを利用することが多い[1]。しかし、フィードバックを多く要求することからユーザへの負担が大きい。フィードバックする際に不適合コンテンツが十分に考慮されていない。

3. 研究目的と研究アプローチ

本研究は、検索対象における知識のないユーザが、複数のコンテンツ獲得を支援することを目的とする。アプローチは以下の3つとする。

3.1. コンテンツベクトルの表現

従来、コンテンツの特徴が十分に表現されず、類似しないコンテンツが同じクラスタに含まれてしまう。そこで、コンテンツの特徴としてタイトル(コンテンツの主な情報)とサマリ(キーワードに近い情報)を利用したコンテンツベクトルを表現する。またコンテンツ全体の特徴値算出には低頻出でも特徴的な語を抽出するために、TF(Term Frequency)値の影響を抑え、低頻出でも特徴的な単語を重視する。

Document Clustering Using User Context Queries
Tomonobu Obeyama, Motoei Azuma, Dept of IMSE, Graduate School of Creative Science. & Engineering, Waseda University.

3.2. ユーザの検索意図把握

従来の適合性フィードバックによるクエリの生成ではフィードバック数が多く必要とされ、ユーザへの負担が大きい。また閲覧したコンテンツだけではユーザの興味を学習する際に偏りが生じやすい。

そこでユーザの検索意図を把握するため、ユーザが必要とする情報が含まれると考えられるコンテンツを収集する。適合・不適合コンテンツと関連コンテンツを利用することで、ユーザの検索意図を考慮した検索システムを実現する。

3.3. クラスタリング結果へのフィードバック

従来のクラスタリング結果にフィードバックする研究において、不適合コンテンツが十分に考慮されておらず、適合クラスタ内に不要なコンテンツが含まれる。そこで本研究では不適合クエリも考慮したフィードバックを行い、コンテンツ収集を支援する。

4. 提案システム

4.1. システム概要

まずユーザは検索キーワードを入力し、コンテンツの適合・不適合判定を行う。次にクエリが提供され、ユーザはクエリを選択する。最後にクエリを利用したクラスタリング結果を提示する。システム概要を図1に示す。

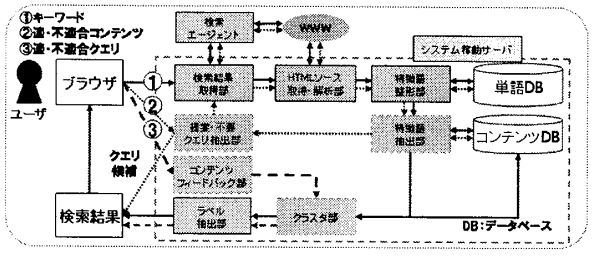

図1. 提案システム概要図

4.2. コンテンツベクトル

従来、コンテンツベクトルは頻出する語の影響を受けやすく、不要なコンテンツが多く含まれていた。

そこで、本研究ではタイトル(title)、サマリ(summary)、全文(plain)を考慮する。さらにTF値の影響を抑え、頻出しないが特徴的な語を抽出する。これによりコンテンツにおいて特徴的で、検索結果閲覧時に類似するコンテンツが同じクラスタに分類されやすい。

コンテンツjのコンテンツベクトル算出式を(1)に示す。

$$d_j = d_j^{plain} + \alpha d_j^{title} + \beta d_j^{summary} \cdots (1)$$

タグ別のコンテンツベクトル算出式を(2)に示す。

$d_j^{tag} = w_t^{j \cdot tag}, \cdots$ tagは$plain, title, summary$をとりうる・・・(2)

ある単語 t のコンテンツ d におけるタグ tag の特徴値算出式を(3)に示す。

$$w_t^{d \cdot tag} = \log\log\log\log(tf_t^{d \cdot tag} + 1.0) + 1.0) + 1.0) + 1.0) \cdot \log(N / df_t) \quad \cdots(3)$$

$w_t^{d \cdot tag}$: 単語 t、コンテンツ d、タグ tag の特徴値

$tf_t^{d \cdot tag}$: 単語 t、コンテンツ d、タグ tag の頻出度

df_t : 単語 t が出現するコンテンツ数

算出式(3)で TF 値の影響を下げ、算出式(1)でタイトルとサマリに出現する単語の特徴値をあげることでタイトルとキーワードに関係した特徴を反映する。

4.3. ユーザの検索意図把握

従来、適合・不適合コンテンツのみを利用しクエリ生成を行っていた。しかし、ユーザにフィードバックを多く要求することから負担が大きい。そこで、ユーザの検索意図を把握するため、関連コンテンツを収集し利用する。具体的には同一クラスタ内コンテンツ、リンク先コンテンツ、同一ドメイン内コンテンツを利用する。

4.4. クエリの生成

適合・不適合コンテンツ、収集した関連コンテンツも利用したクエリ推薦を行う。クエリの算出式(4)を示す。

$$RQ(G,w) = \tilde{G} + \mu\tilde{D} + \nu\tilde{L} \quad \cdots(4)$$

$$\tilde{K} = \frac{\alpha}{|K^+|}\sum_{d_j \in K^+} d_j - \frac{\beta}{|K^-|}\sum_{d_i \in K^-} d_i \quad \cdots(5)$$

$\tilde{K}$はG, L, Dmをとりうる　α=2.0、β=0.5

μ、ν: 関連コンテンツパラメータ

提案クエリベクトル（$RQ(G,w)$）算出には適合コンテンツに関連する集合として以下を利用する。

- G^+ : 適合クラスタ内コンテンツ集合
- L^+ : 適合コンテンツのリンク先コンテンツ集合
- D^+ : 適合コンテンツと同一ドメイン内で検索キーワードにヒットするコンテンツ集合

これらを利用し、提案クエリ ベクトルの特徴値を算出する。一方、不適合コンテンツに関連する集合として、以下を利用する。

- G^- : 不適合クラスタ内コンテンツ集合
- L^- : 不適合コンテンツのリンク先コンテンツ集合
- D^- : 不適合コンテンツと同一ドメイン内で検索キーワードにヒットするコンテンツ集合

これらの特徴値を減じることで、ユーザの検索意図に適した単語の特徴 値のみが高い値となる。クエリベクトルの上位 N 件を提案クエリ、下位 N 件を不要クエリとしユーザに提示する。そして、ユーザが有用と判断したクエリをコンテキストクエリ 集合（CQ(G,w)）、不要と判断したクエリを非コンテキストクエリ集合（NCQ(G,w)）とする。

4.5. ユーザの検索意図推移への対応

クエリの生成を行う際、2 度目以降のクエリ推薦では、ユーザの検索意図が推移した場合には前回と異なるキーワードが重要であると考えられる. そのため、前回推薦したクエリ（$RQ(G,W)_{before}$）と今回（$RQ(G,W)_{new}$）の差分が大きいものを重要と考える。算出式を　(6)に示す.

$$RQ(G,W) = RQ(G,W)_{new} - RQ(G,W)_{before} \quad \cdots(6)$$

4.6. コンテンツへのフィードバック

コンテキストクエリ、非コンテキストクエリを利用し、コンテンツベクトルを更新する。従来では考慮されていない不適合クエリを考慮した補正式を(7)に示す。

$$\tilde{d}_j = d_j + \lambda(CQ(G,w) + NCQ(G,w))$$

$$\lambda = \begin{cases} \lambda \\ \lambda - k((CQ(G,w) \cap \mathrm{d}_j \neq \phi) \\ \cap(NCQ(G,w) \cap \mathrm{d}_j \neq \phi)) \end{cases} \quad \cdots(7)$$

λ, k: 適合・不適合クエリパラメータ

(7)式によりコンテキストクエリのみが含まれるコンテンツが同一クラスタに含まれやすい。

4.7. クラスタリング手順

①コンテンツベクトル間の類似度を計算し、最も高いベクトルを 同一クラスタに含める。② 閾値を下回るまで①を繰り返す 。③残りのコンテンツをその他 に含める。④ラベルのないクラスタに対し、クラスタ内で最も高い特徴値を持つ単語をラベルとしクラスタリングを終了する。

5. 評価実験と考察

本研究のプロトタイプを実装し、評価実験を行った。本システムによるクエリ生成を行い、クラスタリング結果にフィードバックさせたところ、高い適合率を得られた。

6. 終わりに

本研究では、適合性フィードバックを利用したクラスタリング検索システムを提案した。ユーザの検索意図を把握するため、関連コンテンツを利用することで、ユーザの検索意図に適したクエリを推薦し、不適合クエリをフィードバック することで検索システムとしての有効性を示せた。今後の課題として、アルゴリズムの改良がある。

参考文献

[1]江口浩二他:"漸次的に拡張されたクエリを用いた適応的文章クラスタリング法",電子情報通信学会論文誌,D-I Vol.J82-D-I No.1 pp.140-149(1999 年 1 月)