2K-08

視覚的コンテキストとクラスタリングによるWeb検索結果の提示

平田 宗一郎[†] 井桁 正人[‡] 寺田 実[§] 丸山 一貴[¶]
明星大学 情報学部[†¶] 楽天株式会社[‡] 電気通信大学 情報理工学部[§]

1. はじめに

我々は普段知りたい情報を得るために，検索エンジンを利用することが多い．しかし，検索結果の断片的な情報と実際の Web ページとの相違によって，目的の情報に到達することに時間がかかってしまうことがしばしばある．上記を解決する手法として，既存の検索結果を拡張するもの[1][2][3]と，ページにある検索クエリ周辺を画像として切り取って，視覚的コンテキスト(以下，VP という)としてユーザへ提示するもの[4]がある．いずれの手法もページの情報を更に検索結果へと加えることによって，ユーザへページにある情報の手がかりを提供している．ここで，手がかりではなく目的となる情報そのものを検索結果として提示すれば，より早く情報を入手出来ると考えた．また，似た情報ごとにクラスタとすることにより，情報について横断的に調べたい場合と 1 つの事柄を重点的に調べたい場合に対応する．

本研究の目的は検索結果を見るだけで目的の情報を得られるようにすることである．具体的には，丸山ら[4]の手法を 2 点変更することにより検索体験の向上を図る．1 つは VP の切り出し範囲の修正であり，もう 1 つは類似した VP をクラスタリングによりグループ化することである．

2. 提案手法

本提案手法は VP 生成部とクラスタリング部に分けられる．

2.1. VP 生成部

VP 作成までの流れを次に示す．

1. Google 検索から，クエリに対応した候補ページのリストを取得
2. 候補ページをブラウザでレンダリング
3. レンダリングされた各候補ページからクエリに関係する部分を切り出す

本手法ではユーザが検索を行う際，各 Web ページを閲覧せずに検索結果でもクエリに対する情報を詳しく知りたいものであると考え，既存手法よりまとまった単位で情報を提供する．したがってレンダリングされた候補ページからクエリに対する情報を段落や画像といった単位で切り出すために，DOM[1]単位で切り出してユーザへ提示する．VP の例は図 1 に示す．

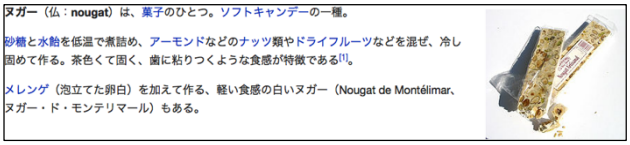

図 1 提案手法の VP

切り出す DOM は，次のアルゴリズムで決められる．

- 候補ページ内から，検索クエリの文字列を自身の text 要素に含んだすべての DOM を探す
- クエリを含んだ DOM の周辺に一定以上の大きさの画像が存在すれば，途中にあるものも含めてクエリ DOM と画像 DOM の全領域が含まれる最小の長方形を VP として切り出す

Search engine result page of clustered snippets with visual context

[†] Soichiro Hirata, School of Information Science, Meisei University
[‡] Masato Igeta, Rakuten Inc.
[§] Minoru Terada, Faculty of Informatics and Engineering, The University of Electro-Communications
[¶] Kazutaka Maruyama, School of Information Science, Meisei University

[1] Document Object Model，ここでは HTML をツリー構造として扱った場合のノード．

- 一定以上や以下の大きさの VP は排除

2.2. クラスタリング部

クラスタリング部では，作成された VP を似た情報ごとにクラスタリングする. VP 同士を比較する指標として，VP を画像として扱う案と，VP 内の文字列を利用する案の 2 つを評価する．画像として扱う案では，SURF 法を用いて局所特徴点を抜き出してクラスタリングした．文字列を利用する案では，文章の特徴ベクトル化に広く使われている TF-IDF 法を用いて行い, その後クラスタリングした.

また，両案の最後に行うクラスタリングでは，共に 2 つのクラスタリングアルゴリズムを用いて評価した．1 つは k-means 法であり，広くクラスタ化に利用されている．しかし，クラスタ数を最初に指定する必要があるため，今回のような最終的なクラスタ数が推測できない問題に対しては適用が難しい．そのため実験ごとに正解集合と同じクラスタ数を指定する．もう 1 つは Affinity Propagation(以下，AfP という)であり，こちらはクラスタ数を事前に指定する必要がない.

クラスタリングを行う際は，Web ページごとではなく，個々の情報である VP ごとに行う．これは 1 つの Web ページは複数の情報のかたまりであるという考えからであり，よって別々の Web ページでも似た情報であれば同じクラスタへ含まれ，同一 Web ページ内の情報であっても，かけ離れた情報であれば別のクラスタへと分離される.

3. 実験

実験では 3 つの VP 群に対して，クラスタリング部で作成された 4 つの結果と，あらかじめ人手で分類した正解集合とを比較した．評価は，広くクラスタの評価方法として利用されているエントロピーと純度の 2 つを用いた．エントロピーは低い方が，純度は高い方がより正解集合に近く，結果が良好であることを示す. 検索クエリを”通天橋”として検索した VP 群に対する, 4 つのクラスタリングの評価結果は，表 1 のとおりである.

表 1 クラスタリングの評価結果

クラスタ方法	エントロピー	純度
画像/k-means	0.7880877	0.6521739
画像/AfP	0.7080904	0.3913043
文字列/k-means	0.6380929	0.6521739
文字列/AfP	0.7045429	0.5217391

4. 考察

実験結果は，全体的に画像によるクラスタリングより文字によるクラスタリングの方が結果は良好であった．画像によるクラスタリングでは，k-means 法が AfP よりも結果が良好である場合は多かった．しかし，AfP が上回った場合もあり，また k-means 法にはクラスタ数指定の問題もある．よって本論文の VP には，文字列として AfP でクラスタリングした方がよいことが考えられる.

5. おわりに

我々は視覚的コンテキストによる，新たなアプローチによる検索結果の掲示方法の可能性を示した．今後はクラスタリングの精度を上げるため，VP を, 文字と画像の両方の情報を用いてクラスタリングすることが考えられる.

参考文献

[1] 西海俊秀,“インテリジェントアイコンによる Web 検索結果閲覧支援,”人工知能学会全国大会論文集, 第 25 巻, pp. 1347-9881, 2011.

[2] 高見真也 , 田中克己,“ウェブ検索結果における検索目的に応じたスニペット生成,”情報処理学会論文誌ジャーナル, Vol.49, No.4, pp. 1648-1656, 2008.

[3] 原島純 , 黒橋禎夫,“PLSI を用いたウェブ検索結果の要約,”言語処理学会第 16 回年次大会論文集, pp. 118-121, 2010.

[4] 丸山一貴, 井桁正人, 寺田実, “視覚的コンテキストによる検索結果提示とナビゲーションの可能性”, 2014-HCI-157(26), pp.1-6, 2014.