4M-8

文書類似度を考慮した検索システムの開発

佐藤　優介†　　吉田　博哉†
神戸情報大学院大学　情報技術研究科†

1.　はじめに

1.1.　研究背景

近年，スマートフォンなどの情報端末の普及に伴いWeb上には膨大な情報が公開されている．これらの中から目的の情報を探し出すには，GoogleやYahoo!，Bingといった検索エンジンを利用するのが一般的である．

Googleが実施した利用者の検索エンジンの利用動向調査の結果，検索結果の上位21件目以降は閲覧される可能性が低下することが明らかになっている．そのため，情報発信者は，自身が作成したWebページが検索結果の上位に表示されるようにSEO対策を行う．しかし，検索目的は情報受信者により異なり，場合によっては不要な情報となりうる．そこで，情報受信者は，検索結果の中から閲覧するページを判断する基準が必要となる．

1.2.　先行研究

検索結果として得られた情報を分類し，利用者（情報受信者）に提示する検索支援を行うための研究は多く行われている．類似研究として，単語グループに基づくWeb文書クラスタリング[1]や，内容類似度に基づくWEBサイト間関連度可視化に関する検討[2]がある．本研究では，検索結果を分類し可視化表示を行う．

1.3.　本研究の目的

本研究は，多くの情報の中から目的の情報発見までの閲覧回数の削減を目的とする．そのため，検索結果を分類し，絞り込み機能を有するシステムの開発を行い閲覧ページの判断を支援する．

2.　システム概要

本システムは，利用者によって入力された検索ワードに対し，検索APIを用いて検索結果を取得する．取得した検索結果を文書類似度で分類する．そして，検索結果を従来の一覧表示と分類結果を可視化した結果を表示する．

Developments of the search system that taking account of document similarity.
† Yusuke Sato
† Hiroya Yoshida
Kobe Institute of Computing(†)

3.　システム詳細

3.1.　画面構成

本システムの画面構成を図1に示す．

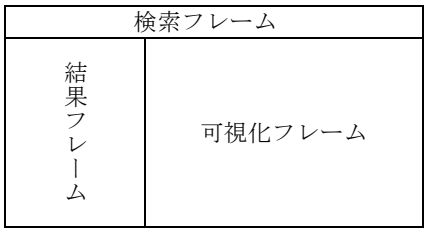

図 1：画面構成

本システムでは，検索フレーム，結果フレーム，可視化フレームといった3つのフレームで構成する．検索フレームには，検索ワード入力用テキストボックスを設置する．また結果フレームには，従来の検索エンジン同様にタイトルと要約を一覧表示する．そして，可視化フレームには，分類結果を表示する．

3.2.　利用フロー

本システムの利用フローを図2に示す．

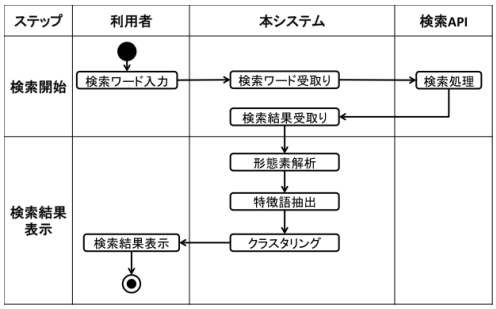

図 2：利用フロー

検索開始ステップでは，検索ワード入力から検索結果受取りまでの処理を行う．本システムで利用者が入力した検索ワードを受取り，検索APIで検索を実行する．その結果URLとページタイトル，要約を取得する．本システムでは取得したページタイトルと要約を1つの文書として扱う．なお，検索APIはBing Search APIを用いる．

検索結果表示ステップでは，形態素解析から検索結果表示までの処理を行う．このステップの詳

細を次節で述べる．

3.3. 処理の詳細

3.3.1. 形態素解析

形態素解析処理では，特徴語抽出のための前処理として文書に出現する単語を抽出する．形態素解析器には，MeCab を用いる．辞書として IPA 辞書に加え，最新の話題語や専門用語に対応するため，はてなキーワードに登録されている単語を辞書に登録する．

3.3.2. 特徴語抽出

特徴語抽出処理では，クラスタリング処理で扱うベクトル削減や後述する可視化フレームで特徴語を表示するために，文書の特徴語を，形態素解析によって抽出された単語から tf-idf 法を用いて算出する．本研究では，tfidf 値が平均以上の単語を特徴語とする．なお，検索ワードに用いられた単語は，特徴語から除外する．

3.3.3. クラスタリング

クラスタリング処理では，検索結果を分類し表示するためにクラスタリング処理を行う．処理速度を優先し，計算量の少ない k-means 法を用いてクラスタリングを行う．なお，k=5 で分類する．

3.4. 検索結果表示

本システムでは，結果フレームに従来の一覧表示を表示し，可視化フレームにクラスタリング結果を表示する．なお，可視化フレームは，BubbleTree[3]を用いて図 3 に示す円で表示する．

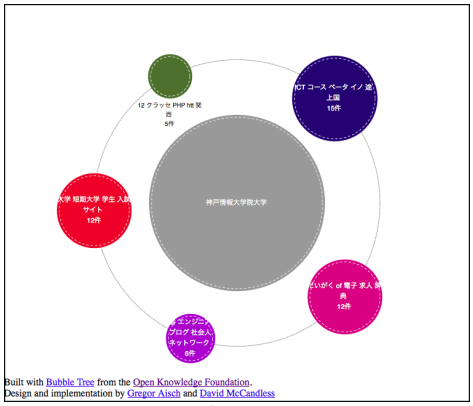

図 3：可視化フレーム

図 3 に示すそれぞれの円は 1 つのクラスタ（分類結果）を示している．各円には，属する Web ページの件数と特徴語を表示している．

3.5. 絞り込み機能

本システムでは，可視化フレームに表示された検索結果をもとに，結果フレームを絞り込む機能を有する．図 3 に示す可視化フレーム内の任意の円を押下すると，図 4 に示すように選択した円に属する Web ページの特徴語一覧が表示される．

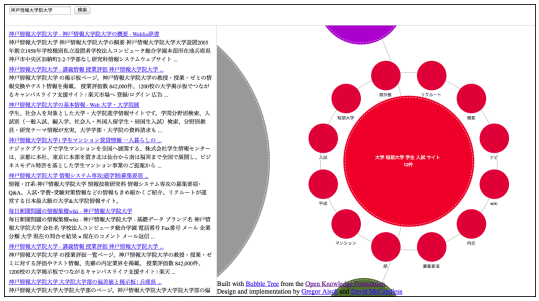

図 4：絞り込み機能

この時，結果フレームの一覧表示も，選択した円に属する Web ページのみが表示される．さらに，可視化フレームにて選択した円の周囲に表示された特徴語を選択することで，より詳細な絞り込みが可能となる．

4. 検証実験

4.1. 方法

被験者には，画像や情報の一部を提示し，本システムを用いて検索させ，本システムでの閲覧回数と Bing を用いて検索結果を上位から順次閲覧したと仮定した閲覧回数を比較する．

4.2. 結果と考察

5 名の被験者により検証をした結果，すべての被験者において閲覧回数削減の結果が得られ，本システムで実装した検索結果の分類，特徴語による絞り込み機能の有用性を示せたと言える．

5. まとめ

本研究では，検索結果を分類した結果を表示し，特徴語による検索結果の絞り込み機能を有するシステムの開発を行った．検証実験の結果，閲覧回数の削減傾向がみられ，一覧表示のみによる検索結果と比べ，本システムの有用性を示した．

今後の課題として，本研究では，クラスタリング手法に非階層的クラスタリングを用いたが，階層的クラスタリングを用いることで，階層毎の詳細な絞り込みが可能になると考える．そのため，クラスタリング手法について比較検証する必要がある．

参考文献

[1] 石川貴浩，奥平雅士 ： 内容類似度に基づく WEB サイト間関連度可視化に関する検討，映像情報メディア学会技術報告，Vol. 35， No. 8， pp. 111-113， 2011.

[2] 仁科朋也，内海彰 ： 単語グループに基づく Web 文書クラスタリング，自然言語処理/言語処理学会編，Vol. 17，No. 4，pp. 23-41，2010.

[3] BubbleTree ： http://okfnlabs.org/bubbletree/