レーダーチャートを用いた柔軟なリランキング手法の実装

荒 深 康 夫[†]　河合 由起子[†]
張 建 偉[†]　熊本 忠彦[††]

A Page Re-Ranking System by a Radar Chart for Flexible Web Information Retrieval

Yasuo Arahuka,[†] Yukiko Kawai,[†] Jianwei Zhang[†]
and Tadahiko Kumamoto [††]

1. はじめに

近年，検索エンジンの結果を利用して，検索結果を分類する研究や，再検索の手間を軽減するインタフェースに関する研究開発が多く行われている[1)2)]，基本的な検索結果の分類手法は，検索キーワードを補完するサブキーワードを出現頻度や共起関係より抽出し，それらをサブキーワードごとに検索結果を分類する手法が一般的である[3)4)]．再検索の手間を軽減する手法では，検索キーワードを入力ボックスにユーザが記入する必要がなく，例えば検索結果で提示されるタイトルとスニペット（概要）に表示されている単語をマウスで選択してやると,「強調（ＡＮＤ検索)」「削除（ＮＯＴ検索)」が容易にできる[5)]．これらの研究開発により，サブキーワードを考慮あるいは入力する手間が軽減できる．しかしながら，これらの再検索手法は，複数のサブキーワードの関係性（優劣）を考慮した検索はできない．

そこで，本研究では，検索結果から複数のサブキーワードを推薦し，それらのサブキーワードの重みをユーザが容易に組み合わせて各サブキーワード間の優劣を決定することで，柔軟な検索が可能なリランキングシステムを提案する．この各サブキーワードの重みの容易な指定方法として，レーダーチャートをインタフェースとして利用する.(図 1) 具体的には，まず，検索エンジンのＡＰＩを用いて検索結果を取得する．取得した検索結果から $tf \cdot idf$ 値とタイトル長を考慮しキーワード上位 M 個の単語をサブキーワードとして抽出する．この抽出されたサブキーワードがレーダーチャートの各項目となり，ユーザへ推薦される．また，

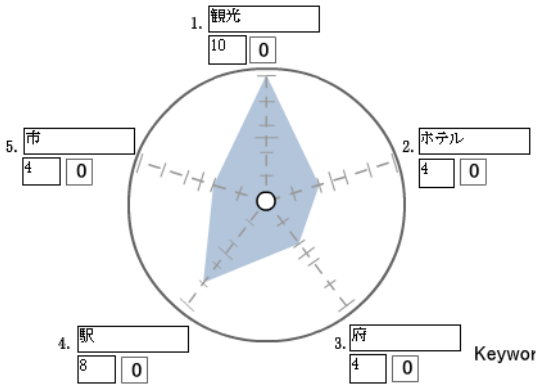

図 1　レーダーチャートによる優劣を考慮したリランキングの一例

そのキーワードの特徴値をチャート値として提示する．ユーザはこのレーダーチャートの各項目のチャート値を自由に変更することで，複数のサブキーワード間の重みを組み合わせることができ，その結果優先順位を考慮したリランキングが実現できる．本論文では，レーダーチャート検索システムについて述べるとともに，100 名による評価実験より項目抽出やインタフェースについて検証し考察する．

2. システム設計

2.1 システムの概要

提案システムは，検索結果から複数のサブキーワードを自動抽出し，ユーザがそれらのサブキーワードの重みを自由に設定することで，効果的なリランキングができる．図 1 に提案システムの概要を示す．まず，サーバはユーザの入力した検索キーワードを受信すると，yahoo!の検索 API を用いて検索結果を取得する．次に，取得した検索結果の各ページのタイトルと概要

† 京都産業大学
Kyoto Sangyo University
†† 千葉工業大学
Chiba Institute of Technology

であるスニペットを取得し，形態素解析する．さらに，各単語の出現頻度となる $tf \cdot idf$ 値を算出し，その値とタイトル・スニペット長を基に検索キーワード以外の上位 M 個の単語をサブキーワードとして抽出する．この抽出したサブキーワードと値をレーダーチャートの項目とそのチャート値としてユーザへ提示する．(図 1)

次に，ユーザは提示された各項目（サブキーワード）のチャート値を 10 段階で評価する．この時，ユーザは項目のキーワードを新たなキーワードへ書き換えることもできる．サーバはユーザから各項目とそれらのチャート値を受信すると,「各項目に入力されたチャート値を要素とするベクトル」と「各ページのサブキーワードの $tf \cdot idf$ 値を要素とするベクトル」との cos 相関値を算出する．この cos 相関値を基にリランキングした結果をユーザへ提示する．

例えば，ユーザが検索キーワード「首相」を問い合わせた結果,「民主党」「与党」「野党」「外交」「自衛隊」の項目が抽出され，チャート値が (0．3， 0，3， 0．2， 0．1， 0．1）であったとする．この時，ユーザが「自衛隊」と「野党」に興味があった場合，該当項目のチャート値を「0．9」と「0．7」へと高めに変更すると，(0．3， 0．3， 0．9， 0．1， 0．7）の値をサーバは受信する．そして，それらの値を基に「首相」の検索結果から「自衛隊」＞「野党」＞「民主党」,「与党」＞「外交」を基準としたリランキング結果を得ることができる．

2.2 レーダーチャート項目とチャート値の抽出

yahoo! API を用いて検索結果上位 N 件を取得する．次に，取得した検索結果から，実時間処理を考慮し，各ページのタイトルと概要であるスニペットを取得する．さらに，取得した情報を形態素解析し，名詞のうち,「名詞-非自立-一般」と「名詞-接尾-一般名詞-接尾-人名」を除く品詞の単語に該当するものを抽出する．対象とした品詞はこれまでの実験より選定した．各ページの抽出されたそれらの単語から，各単語の $tf \cdot idf$ 値を下記より算出する．

$$tf \cdot idf = \frac{(T_i\text{中の単語 } j \text{ の出現総数})}{(T_i\text{中の総単語種数 N})} \times log\frac{(\text{検索件数 M}+1)}{\text{単語 } j \text{ が出現するページ数}} \quad (1)$$

ここで，事前実験よりタイトルはスニペットと比較すると特徴的な単語が多いた，4 倍の重みとした．また，取得したタイトルやスニペットは 10 単語程度と短いものや，全く単語のないものの影響を考慮し，単語 j のチャート値を下記より算出した．

$$j \text{ の } tf \cdot idf \text{ 合計} \times \frac{(j \text{ の出現総数})}{(\text{総ページの総単語数})} \quad (2)$$

レーダーチャート項目となるサブキーワードは，検索キーワード以外の式（2）の上位 M 単語とする．

2.3 チャート値によるリランキング法

ユーザは提示された各項目のチャート値を 0〜10 の整数値で評価する．チャート値への変換は，各単語の $tf \cdot idf$ 値を元に以下の式を基に換算する．

図 2 チャート値を変更後の検索結果

$$\begin{aligned} y &= 0 \quad (x = 0) \\ y &= tf \cdot idf \text{ 値の最小値} \quad (x = 1) \\ y &= \frac{(tf \cdot idf \text{ 値の最大値} - tf \cdot idf \text{ 値の最小値})}{9} \\ &\quad \times (x-1) + tf \cdot idf \text{ 値の最小値} \quad (1 < x < 10) \\ y &= tf \cdot idf\text{値の最大値} \quad (x = 10) \end{aligned}$$

x は 0〜10 の整数値である．具体的には，任意のキーワードの $tf \cdot idf$ 値の最大値が 0．09 で最小値が 0．03 とする．この時の換算式は

$$y = (0.09 - 0.03)/9 \times (x-1) + 0.03 \ (1 < x < 10)$$

となり，y の値をレーダーチャートのチャート値として提示する．

この時，ユーザは提示された項目を別の新たなキーワードへ書き換えることができる．システムは，ユーザから 5 つの項目のキーワードとそれらのチャート値を受信すると，各項目のチャート値をベクトルの要素とするベクトル V_q=(v_{q1}， v_{q2}， v_{q3}， v_{q4}， v_{q5}) を決定する．

次に，検索結果から取得した各ページから，項目の単語をベクトルの要素とし，それらの単語の $tf \cdot idf$ 値を要素値とするベクトル V_p=(v_{p1}， v_{p2}， v_{p3}， v_{p4}， v_{p5})，(p=1,...， N，N は取得した総ページ数，また，その単語が存在しない場合，値は 0) を決定する．以上のベクトル V_q と V_p の相関を下記より算出する．

$$sim(V_q, V_p) = \frac{(v_{q1}v_{p1} + \cdots + v_{q5}v_{p5})}{(\sqrt{{v_{q1}}^2 + \cdots + {v_{q5}}^2} + \sqrt{{v_{p1}}^2 + \cdots + {v_{p5}}^2})} \quad (3)$$

この cos 相関値の高い順にランキングした結果をユーザへ提示する．つまり，レーダーチャート値を 10 にしたときは，そのサブキーワードに対する $tf \cdot idf$ 値が最も高いページが上位にランキングされるようになる．

3. 評 価 実 験

本章では，構築したレーダーチャート検索システムを検証する．

レーダーチャート検索システムは，PHP Version

5．2．6，FLASH ver8．0 で開発し，検索エンジンには，Yahoo!検索 WebAPI Ver．1．0，形態素解析には MeCab Ver．0．97 を用いた．システムの項目抽出およびリランキングに関する評価は，全国の 20〜60 歳の男女 100 名を対象に公開している本システム[7]を利用してもらい，その結果を検証した．

検索結果数 N は，これまでは 50 件で行っていたが，事前実験より，50 件，100 件，150 件，200 件で検証した．検討結果より，10 秒以内となる 150 件とし，上位 150 件全ページのタイトルとスニペットを取得した．尚，リランキングに要する時間はリランキング前の検索時間と同程度で瞬時である．

実験では指定の検索キーワードの検索結果からレーダーチャートを用いてリランキングした結果 150 件から指定のページを探す．

また，レーダーチャート項目数は 5 単語として提示したが，本システムの特徴としてユーザが自由に変更できる点を利用し，項目外に上位 15 単語（項目 5 単語を含む）を関連語として推薦・提示した．以下では，インタフェース，レーダーチャート項目ならびに関連語として提示した 15 単語とリランキングについての検証を行う．

3.1 インタフェース

構築したシステムでは，ユーザが検索キーワードを入力すると，最初は既存の検索エンジン（Yahoo!）の検索結果を得ると同時に，レーダーチャートが提示されており，ユーザは項目から検索キーワードに対するサブキーワードを確認できる．また，チャート値より，それらのサブキーワードの優劣も把握できる．

次に，ユーザは提示されたサブキーワードから興味のあるキーワードを複数選択し，優劣を考慮してチャート値を変更できる．図 2 は，「戦争」＞「支援」＞「情勢」と変更後にリランキングされた結果である．また，各項目に興味のあるキーワードがなければ，ユーザ自身で項目を容易に書き換えることもできる．

3.2 項目と関連語抽出に関する考察

3.2.1 項目と関連語抽出に関する定性的評価

提案システムでは，ユーザが検索キーワードを入力すると，検索結果から式（2）を用いてレーダーチャートの項目と関連語となる 15 単語を抽出し，ユーザへ提示する．本節では，提示されたレーダーチャートの項目と関連語に対する満足度に関する定性的評価を行う．

満足度を評価するために，被験者が入力した検索キーワードに対してシステムが提示したレーダーチャートの項目 5 つに対して，被験者に 5 段階評価（適切，どちらかといえば適切，どちらかといえば不適切，不適切，覚えていない／分からない）をしてもらった．また，関連語 15 単語（項目 5 単語を含む）に対して，5 段階評価（役に立った，どちらかといえば役に立った，どちらかといえば役に立たなかった，役に立たなかった，覚えていない／分からない）をしてもらった．図 3 に被験者の評価結果を示す．評価結果よりレーダーチャート項目として提示した 5 項目に関しては，「覚えてない／分からない」を除いた場合，約 60 %の割合で「適切」又は「どちらかといえば適切」となった．また，関連語では「覚えてない／分からない」を除いて，70 %が役に立ったと回答した．

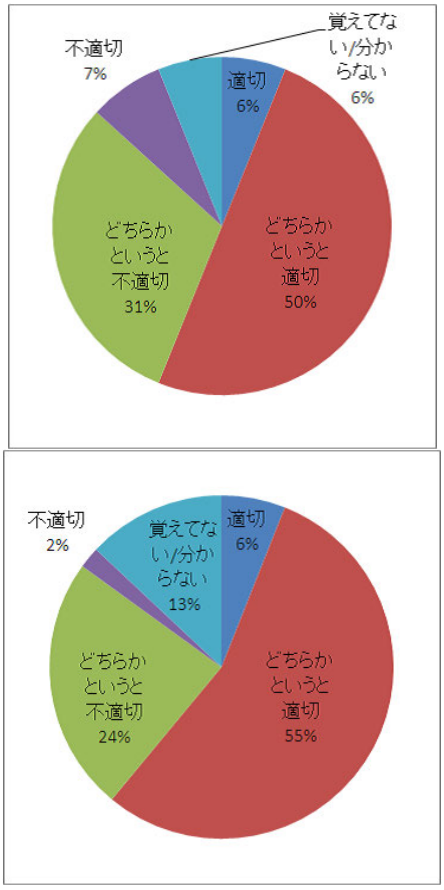

図 3 項目 5 単語（上）と関連語 15 単語（下）に対する満足度評価実験結果

3.2.2 項目と関連語抽出に関する定量的評価

図 4 にレーダーチャートによるリランキングの際に利用した項目の上位 9 個を各検索キーワードごとに示す．青色グラフが本システムが提示した 5 つの項目である．その他の白色グラフは，リランキングの際にユーザが提示した項目名を削除して新たに記入したものである．図より，システムが提示した項目の利用頻度が高いことが確認できる．

さらに，各項目のチャート値の入力では，最初の課題では 19 名の約 2 割がチャート値に優劣をつけた検索を行っており，最終課題の 3 回目には 31 名の 3 割が優劣をつけた検索を行っていた．

以上より，項目と関連語の抽出はリランキングにおいて妥当であることが確認でき，項目抽出ではその有効性を確認することが出来た．

3.3 リランキングに関する考察

3.3.1 達成度に関する考察

被験者に 5 段階評価（すぐに見つかった，どちらかといえばすぐに見つかった，どちらかといえば時間がかかった，時間がかかった，時間がかかり，見つからずにあきらめた）してもらった．図 5 に被験者の評価結果を示す．

グラフより，本システムの達成度が，約 2〜2.5 倍に増加したことが明らかとなった．

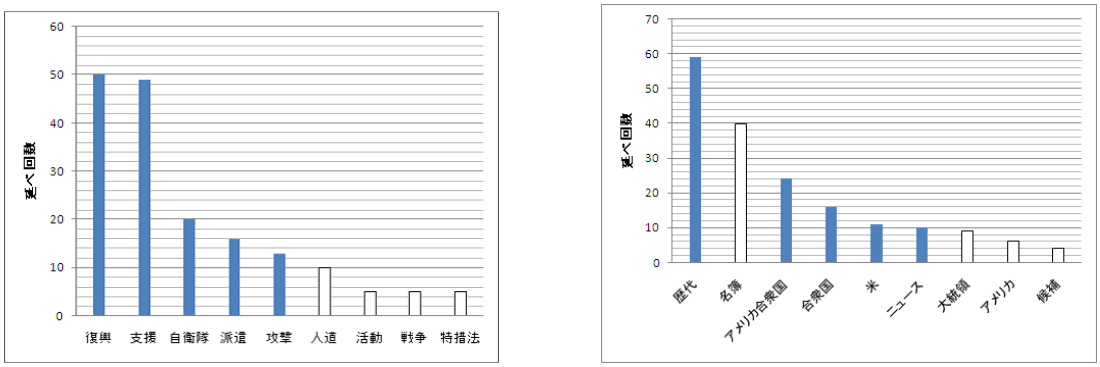

図 4　リランキングの際に利用したチャートの項目名と利用回数

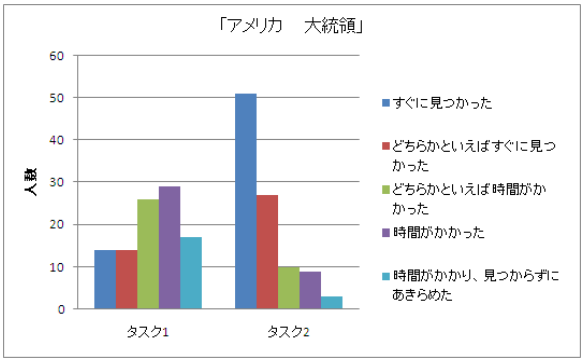

図 5　「アメリカ　大統領」の検索結果に対する達成度評価結果

3.4　今後の課題

実験より，レーダーチャートを用いることで，サブキーワードを再考慮する必要がなく，また，同時に優劣まで設定でき，効果的な再検索ができることが確認できた．

実験に伴い，項目数と提示する検索結果数に関するアンケート調査を行った．普段検索の際に入力するキーワードは 90 %の人が 2 つ以内であった．また，約 80 %の人が上位 30 件までしか確認しないという結果であった．この結果より，今回の 5 項目という項目数は，多い提示であったとも言える．しかしながら，前節の実験より関連語に関しては適切であり項目名の候補となり得るため，今後はレーダチャートの項目数を減らし，関連語として提示する予定である．

検索結果数に関しては，リアルタイム性を考慮し 10 秒程度となる 150 件を対象とした．現在，チャート項目や関連語の抽出対象となる件数は 150 件のままで検索対象となる件数を増やすことで，検索速度を同程度にしつつ 1000 件にも対応できるよう改善を行っている．また，関連語以外にも関連の浅い語の抽出と提示も検討をしており，今後，比較検証を行う予定である．

4.　ま　と　め

本論文では，検索結果から検索キーワードと関連する複数のサブキーワードを自動抽出提示し，それらのサブキーワードの重みをユーザが容易に組み合わせて設定することが可能なレーダーチャート検索システムを提案・改良し，評価実験を行った．

実験より，7 割のユーザが抽出した項目と関連語が適切であったとし，その利用頻度も高く，サブキーワードとしての有用性が確認できた．また，項目のチャート値に優劣をつけることでより効率的に再検索ができていることも確認できた．今後は，検索時間を軽減しつつ対象検索数を増やし，比較実験を行う予定である．また，関連が浅い語の抽出と検証も行う予定である．

謝辞

参　考　文　献

1) H. Zeng, Q. He, Z. Chen, W. Ma and J. Ma: “ Learning to cluster web search results ”, Proc of SIGIR2007. pp. 210-217 (2004)
2) 関 隆宏，和多 太樹，山田 泰寛，廣川 佐千男，検索支援と分析のための多面的検索システム，電子情報通信学会 第 19 回データ工学ワークショップ (DEWS2007)， E1-2， (2007)
3) 野田 武史，大島 裕明，手塚 太郎，小山 聡，田中 克己，Web 検索結果のクラスタリングに用いる話題語の質問キーワードからの自動抽出，電子情報通信学会 第 18 回データ工学ワークショップ (DEWS2006)， 2C-i8， (2006)
4) Clusty the Clustering Engine http://clusty.jp/
5) Rerank. jp http://rerank.jp/
6) 吉田 大我， 小山 聡， 中村 聡史， 田中 克己，Web 検索結果におけるキーワード出現相関の可視化と対話的な質問変換，電子情報通信学会第 18 回データ工学ワークショップ (DEWS2007)， C7-2， (2007)
7) レーダーチャート検索システム http://klab.kyoto-su.ac.jp/%7Enogawa/