4T－4

属性分類による高効率検索方式の検討 †
－検索結果分類キーワードの生成方式－

中川 香織　林 憲亨　新井 克也
NTT ソフトウェア研究所

1．はじめに

現在、インターネット上の情報を検索する手段として、多くのロボット型検索サービスが提供されている。一般に、これらのサービスでは、１語のキーワードに対し１万件以上の検索結果を提示する場合もあるため、その検索結果の中から、目的情報を探し出すのは困難である。このような場合、ユーザは最初に得られた検索結果（一次検索結果）の中から、意味的にまとまりのある小さな検索結果集合（部分検索結果）を抽出し、その単位で目的情報の有無を判断しながら検索を進めていく。部分検索結果を得るためには、二次キーワードを追加する必要があるが、一般に、目的の情報を含む適切な大きさの部分検索結果を得る二次キーワードを発見することは難しい。そのため、最近では、二次キーワードの発見を支援する機能を持つロボット型検索サービスも提供されている。

例えば、シソーラス辞書などから、ユーザが入力したキーワードに関係が深い二次キーワード候補を提示するもの[1]や、一次検索結果中に多く出現する単語を二次キーワード候補として提案するもの[2]がある。しかし、これらの方式は、二次キーワードによって抽出される部分検索結果の性質を考慮していない。そのため、ある部分検索結果が、一次検索結果と同じであったり、異なる二次キーワードに対応する部分検索結果が同じである場合があるという問題があった。

本稿では、これらの問題を解決するために、二次キーワードを生成する際、二次キーワード候補それぞれに対応する部分検索結果の重なり合いを調べ、適切な二次キーワードだけを取得する、新しいキーワードの生成方式を提案する。

2．検索結果を分類する二次キーワードの要件

上記の問題を解決するためには、一次検索結果を互いに重なり合いの少ない部分検索結果に分類できれば良い。

しかし、重なりを少なくするために一次検索結果を細かく分類すると、検索効率が悪くなってしまう。反対に、分類する数が少なければ目的情報の有無を判断する回数も減少するが、部分検索結果が大きくなるため、その中に目的とする情報が含まれているのか判断することが難しくなる。

以上の考察から、検索結果を分類する二次キーワードの要件は、「その中に目的とする情報が含まれているか判断できるよう意味的まとまりを持っており、尚且つ、ある程度の検索結果数を持つ部分検索結果を生成する」キーワードでなければならない。このような二次キーワードを検索結果分類キーワードと呼ぶことにする。

以下では、検索結果の重なり合い（被覆率）に基づいて、検索結果分類キーワード候補の中から適切な検索結果分類キーワードを生成する方式について述べる。

3．検索結果分類キーワードの生成方式

3．1検索結果分類キーワード候補の抽出

検索結果分類キーワードは他との判別が可能である部分検索結果を生成するキーワードでなければならない。そこで、Web ページの内容を端的に表しているタイトルに注目し、ブラウザに表示された時タイトルとなる言葉（一番初めに出てくる言葉）とタイトルタグで挟まれている言葉を検索結果分類キーワード候補として利用する。タイトルタグで挟まれている言葉を利用するのは、HTML の記述が不完全で初めに出てくる言葉が上手く切り出せない時や、タイトルが画像ファイルなどで表されている場合などを考慮したためである。得られたタイトルを品詞分解し、キーワードとなり得る品詞（名詞、動詞、形容詞、形容動詞）を抽出する。

3．2 検索結果分類キーワードの選出

抽出されたキーワードで生成される部分検索結果が一次検索結果を多く含んでいる場合、そのキーワードでグループ化をしても一次検索結果とあまり変わらない。従って、一次検索結果に対する被覆率が決められた数値より高いキーワードは削除する。

---

† A Proposal for Classifying Retrieval Results
Kaori Nakagawa, Noriyuki Hayashi, Katsuya Arai
NTT Software Laboratories

3．3 検索結果分類キーワードの統合

キーワード候補で生成された部分検索結果同士の被覆率を調べる。双方の被覆率が高い場合、お互い同じ検索結果を多く含んでいるので検索結果を統合する。（図１－１）

また、一方の被覆率が高く、もう一方の被覆率が低い場合、被覆率の高い部分検索結果は被覆率の低い部分検索結果の中に含まれている。従って、被覆率の高い部分検索結果はもう一方の部分検索結果に含まれ、その部分検索結果のサブグループとなる。（図１－２）

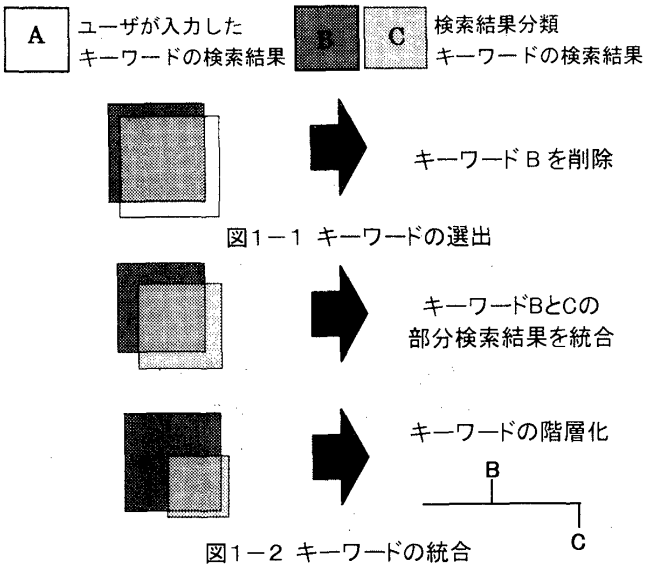

図1－1 キーワードの選出

図1－2 キーワードの統合

4．実験および評価

以上で述べた検索結果分類キーワードの生成方式の効果を確認するために以下の実験を行なった。検索エンジンは goo を利用し、ユーザが入力するキーワードとして「ケーキ」「ランドマークタワー」「広末涼子」を用いて、上で述べた手順に従って検索結果分類キーワードを生成した。この時、一次検索結果に対する被覆率が８０%以上の部分検索結果を生成するキーワードは削除、お互いの被覆率が７０%以上の部分検索結果を生成するキーワードは統合、一方の被覆率が７０%以上、もう一方が１０%以下の部分検索結果を生成するキーワードは階層化（サブグループの生成）を行なった。

表１は「ランドマークタワー」の一次検索結果から抽出した４４のキーワードを対象として、部分検索結果同士の被覆率（ｂ群の部分検索結果／ａ群の部分検索結果）を表したものである。

一次検索結果に対し、被覆率が８０%以上のものは「横浜」だけであり、どの部分検索結果中にも多く存在していることから追加するキーワードとして適切でないことが分かる。キーワードの階層化は多く見られ（図２）、「ランドマークタワー」以外のキーワードでも多く生成できた。

b ＼ a	gallery	市	帆船	本丸	みなとみらい	横浜	・
gallery	100%	6%	0%	8%	4%	3%	・
市	7%	100%	52%	58%	69%	46%	・
帆船	0%	3%	100%	72%	5%	3%	・
本丸	4%	2%	56%	100%	4%	2%	・
みなとみらい	44%	53%	78%	72%	100%	37%	・
横浜	86%	94%	98%	97%	97%	100%	・
・	・	・	・	・	・	・	・

表1：部分検索結果同士の被覆率

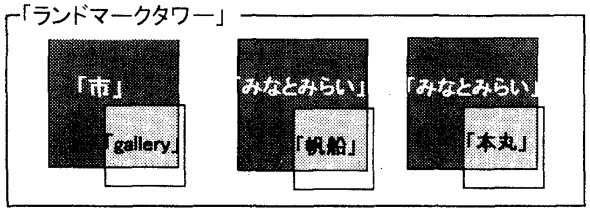

図2：サブグループの生成

得られた部分検索結果の多くは意味的まとまりを持っていた。特に意味がはっきりしていたのは「ケーキ」から生成された「浄水」というキーワードだった。「浄水」の部分検索結果はろ過の道具などのWeb ページの集合であり、他の部分検索結果との違いがはっきりとしていた。

今回はタイトル部分からキーワード候補を抽出するという単純な方法を用いたが、この方法でも意味的にまとまりのある部分検索結果をいくつか抽出することができた。また、被覆率を考慮することで、追加するキーワードとして適切でないキーワードを削除することができる見通しを得た。

5．まとめ

キーワード追加の支援として、検索結果を分類するキーワードをユーザに提示するため、抽出したキーワード候補を検索結果の被覆率に基づいて選出・統合した。その結果、タイトルから抽出したいくつかのキーワードから意味的まとまりのある部分検索結果が抽出され、追加するキーワードとして適切でないキーワードも、検索結果の被覆率を利用することで削除したり、統合することができた。

今後はより多くのサンプル（キーワード）で実験を行い、検索結果分類キーワードの選出・統合のパラメーターの最適化、キーワード抽出方法の有効性の検証を行なう。

参考文献

[1] http://search.kcs.ne.jp/the

[2] 井上他：絞り込み検索語候補の抽出に関する一検討、第 56 回情処全大会、分冊 3,3-95,1998