3F4-3

サーチエンジンクエリ分析による情報タイプの抽出：Web検索利用者の情報要求に即したWeb情報空間の再構成に向けて

Extraction of Information Category Names from Web Search Queries

中渡瀬 秀一[*1]　Hidekazu NAKAWATASE
大山 敬三[*1*2]　Keizo OYAMA

[*1]国立情報学研究所　National Institute of Informatics
[*2]総合研究大学院大学　Graduate University for Advanced Studies

In this paper, we propose a new approach to extract type names which are subcategory of information from web search query logs. In the past research of the query log analysis, related terms and topic words are extracted using the word co-occurrence frequency in query logs. In our approach, we do not analyze the word frequency, but relations between 2 words in queries. The feature of our method is to extract a headword that is related to many modifiers in queries.

1. はじめに

ウェブ上には動画情報・レシピ情報・価格情報といった様々なタイプの情報が蓄積されている．これらの情報が検索される際には，その情報タイプに応じて特有のデータ処理や問い合わせ方法が用いられる．例えば文書検索の処理ではキーワードによる問い合わせに対してその語が含まれる文書のサマリを提示する．一方，動画検索であればコンテンツに検索対象の動画ファイルが含まれるものだけを抽出し，サムネイル形式で提示することが多い．またレシピ情報検索であれば，Google の Recipes 検索に見られるように調理時間や食材という観点での絞込み機能が有用である．さらに近年では特定タイプの情報を効果的に蓄積する仕組みとして分野に特化した CGM[*1]の形成も目立っており，動画投稿サイトやレシピサイトなどはこの代表例である．

このような情報タイプに適した検索処理の開発や情報の効率的蓄積を促すことで Web 検索利用者の情報要求に即した Web 情報空間の再構成が進み，利便性がより向上していくものと考えられる．そのためにも利用者が求めている主要な情報タイプを獲得することは重要な課題であるといえる．

そこで我々はこの情報タイプという語彙領域に注目し，利用者行動が反映された検索サービスログ中のクエリを用いて情報タイプ辞書を作成することを考える．そのために我々は 2 単語からなるクエリを基本とした**二項クエリ構造モデル**を提案している．本稿ではこのモデルとそれに基づいた情報タイプ抽出方法について説明し，実際のクエリログデータを用いて行った情報タイプ抽出実験の結果を報告する．このモデルでは情報検索利用者の要求物が「事物の情報」であると考える．そしてこれを表現するクエリの下位範疇に個々の具体的なクエリが属していると考えるものである．

これまでにも大規模な検索サービスのログから語のグループを抽出する研究として検索語のクラスタリングに関するものが行われている．ログデータの入手が難しいことから公開されている成果は少ないが，例えば [Wen 02] では百科事典サイトの利用者クエリログから得られる約 20,000 語をクエリ内共起や閲覧 Web ページとの共起を用いてクラスタリングしている．[Arita 07] でも検索サイトのクエリログから同様の共起を用いたクラスタリングと階層化によってシソーラス構築を行っている．また [Yasukawa 07] は公開されているキーワード分析ツール[*2]から検索クエリを取得して単語の階層型ベイズクラスタリングを行っている．

これらの検索ログを用いた単語クラスタリング研究と本研究との相違点は，本論文で扱うそれぞれの情報タイプがクラスタリング対象（類似したものの集合）とは異なる点にある．これは「画像」や「価格」といった情報タイプが共起単語による類似度からはクラスタリングされないことから分かる[*3]．この点で本研究は上述の研究とは異なりクエリログから新たな価値を見いだそうとする試みである．

2. 二項クエリ構造モデル

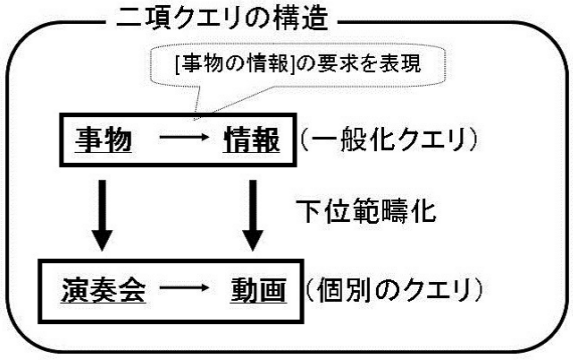

図 1: 二項クエリ構造モデル

ここでは利用者が検索エンジンに要求するものを一般化して「事物の情報」と考える．さらにこれをクエリとして表現するときには具体的な「O T」（O=事物の下位範疇，T=情報の下位範疇）という 2 項形式でクエリにすると考える．このと

連絡先: 国立情報学研究所，〒 101-8430 東京都一ツ橋 2-1-2

∗1　Consumer Generated Media

∗2　http://inventory.jp.overture.com/

∗3　「画像」に対してクラスタリングされやすいものに有名人の名前などがある．

き逆形式の「T O」となる可能性は大規模ログデータを用いた [Nakawatase 11] の調査により低いことが確かめられている．

図 1 に二項クエリ「事物 情報」の下位クエリ「演奏会 動画」の例を示す．個別のクエリ「演奏会 動画」は「事物 情報」の構造を継承し，クエリの各項はそれぞれ「事物」と「情報」の下位語である「演奏会」と「動画」になる．特殊な場合として，事物に対する任意の情報を要求する場合には第 2 項を省略し単一の項からなるクエリとすることもできる．また各項が構造を持つことも可能である．例えば第 1 項が構造を持つ場合には「新春 演奏会 動画」のようなクエリが考えられ，この場合には第 1 項が「新春 演奏会」の 2 語，第 2 項が「動画」となる二項クエリと解釈される*4．

3. 情報タイプの抽出方法

このモデルは個々のクエリにおいて第 2 項が情報の下位範疇，つまり情報のタイプになることを意味している．したがって主要な情報タイプを抽出する際には，検索ログのクエリ集合から十分な頻度を持つ第 2 項の語を収集すればよいことになる．

4. 情報タイプ抽出実験

4.1 実験方法

二項クエリ構造モデルに基づいた情報タイプ抽出の有効性を確認するために実験を行った．クエリログデータはインターネット視聴率調査会社によって作成された大規模なパネル視聴データの 3 セット（2007 年 4 月分，2008 年 11 月分，2010 年 6 月分）*5を用いて作成した．具体的にはパネル視聴データ中の Google（ウェブ検索）に対するリクエストに含まれる検索語（同じクエリの重複を除く）をクエリログデータとしている．さらにこの実験では 2 単語で構成されるクエリだけを対象とした．2 単語クエリはそれぞれが第 1 項と第 2 項に対応する可能性が高く，また検索エンジンに入力される検索語数は 1 語または 2 語である場合が高い比率を占める [Jansen98] ためである．表 1 に実験に用いた各データセットのクエリ数を示す．なおクエリに出現する文字コードの全角英数字・空白は半角英数字・空白に統一して処理を行った．抽出の手順は次の 3 ステップからなる．

1. クエリ集合からそれぞれの第 2 項の語（本実験ではクエリ中の 2 番目の単語）を切り出す．
2. 各語の出現度数をカウントしクエリ数全体に占めるシェアを計算する．
3. シェアが上位となる語を情報タイプとして取り出す．

上記の手順に従って，測定時期の異なる 3 セットのデータに加えて，さらにそのうちの 2010 年 6 月分については重複する内容のクエリを除去しないデータについても抽出処理を行った．このデータでは各クエリに対して検索要求者数によるウェイトが付加されていると考えることができる．これらのデータからの抽出結果を次節以降で比較する．

4.2 実験結果

抽出された語の例として実験結果においてシェアが上位であった語（2010 年 6 月分）を表 2 に示す．表中には獲得された多くの情報タイプ名が見られる．しかし抽出結果中には形容詞，形容動詞，固有名，動詞の語幹など情報タイプ名を表すと限らない品詞類も含まれていたため，これらを除外し得られた情報タイプの経年変化（2007 年，2008 年，2010 年）を表 3 に示す．

表 1: クエリログデータサイズ

測定時期	クエリ数
2007/4	13,647
2008/11	69,412
2010/6	127,733

表 2: 抽出結果（2010 年 6 月分）

順位	抽出された単語
1	動画
2	レシピ
3	ブログ
4	意味
5	歌詞
6	攻略＊
7	画像
8	wiki
9	通販＊
10	口コミ
11	価格
12	評判
13	作り方
14	ランキング
15	地図

＊：情報タイプでない単語

4.3 評価

提案手法における情報タイプ抽出精度を評価するために抽出単語の上位 N 件に含まれる情報タイプ名の比率を作業者が判定して集計した ($N \leq 100$)．ただし後述するように情報タイプになりにくい品詞類についてはここでもあらかじめ形式的に除外している．2007 年，2008 年，2010 年のデータにおけるそれぞれの抽出精度を表 4 と図 2 に示す．

5. 考察

実験結果により，本手法では「動画」「画像」といった情報タイプが正しく抽出されていることが分かる．表 2 において情報タイプでない語は「通販」「攻略」のみであった．これら動詞の語幹由来の名詞は行為や状態変化を表現していることが多い．その場合には情報タイプのような**型**の呼称には適していない．しかしこれらを除外することは容易に可能である*6．

次に主要な情報タイプの経年変化を比較する．表 3 において 2007 年，2008 年，2010 年のデータから抽出された上位 10 件の語を比較すると 3 年間に共通する情報タイプは「動画」，「レシピ」,「ブログ」,「歌詞」,「画像」,「wiki」の 6 語である．

*4 第 2 項が構造を持つ場合も同様

*5 Nielsen Online NetView ローデータ

*6 タイプ名になりうる場合の判断基準を定めることは今後の課題である

表 3: 得られた情報タイプ（経年変化）

順位	2007 年	2008 年	2010 年
1	**wiki**	**レシピ**	**動画**
2	**レシピ**	**ブログ**	**レシピ**
3	**画像**	**wiki**	**ブログ**
4	**動画**	**画像**	意味
5	写真	**動画**	**歌詞**
6	**歌詞**	**歌詞**	**画像**
7	作り方	価格	**wiki**
8	**ブログ**	映画	口コミ
9	壁紙	意味	価格
１０	映画	使い方	評判

動詞，形容詞，固有名などを除外済み
太字は 3 年間で共通している情報タイプ

表 4: 上位 N 件抽出精度（経年変化）

	抽出精度（%）		
上位 N 件	2007 年	2008 年	2010 年
２０	80.0	90.0	90.0
４０	57.5	80.0	77.5
６０	58.3	76.7	76.7
８０	53.8	70.0	72.5
１００	51.0	64.0	67.0

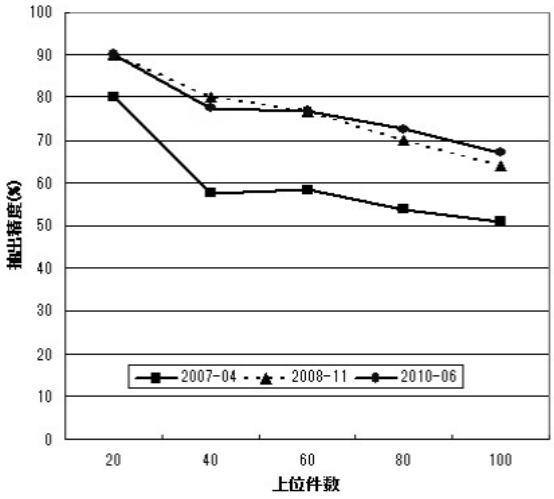

図 2: 抽出精度（経年変化）

最近の 2 年間については，さらに「価格」と「意味」を加えた 8 語が共通していることが分かる．このように主要な情報タイプは経年変化が少ないので，そられに特化した処理技術やサービスの寿命も長期間になることが見込まれる．

抽出精度に関しては上述したように実験により抽出された単語から一部の品詞類*7を除外してから計算したところ上位 100

*7 情報タイプ名とならない他の品詞類として固有名や形容詞，形容動詞，副詞を除外

表 5: 得られた情報タイプ (検索要求者数ウェイト有無の比較)

順位	ウェイトなし	ウェイトあり
1	動画	レシピ
2	レシピ	動画
▷ 3	ブログ	ブログ
4	意味	歌詞
5	歌詞	意味
6	画像	画像
▷ 7	wiki	wiki
8	口コミ	価格
9	価格	ランキング
１０	評判	口コミ
１１	作り方	作り方
▷ １２	ランキング	評判
１３	地図	店舗＊
１４	2ch	使い方
１５	使い方	地図

＊：情報タイプでない単語

表 6: 上位 N 件抽出精度（検索要求者数ウェイト付きとの比較）

	抽出精度（%）	
上位 N 件	ウェイトあり	ウェイトなし
２０	90.0	90.0
４０	77.5	77.5
６０	73.3	76.7
８０	68.6	72.5
１００	67.0	67.0

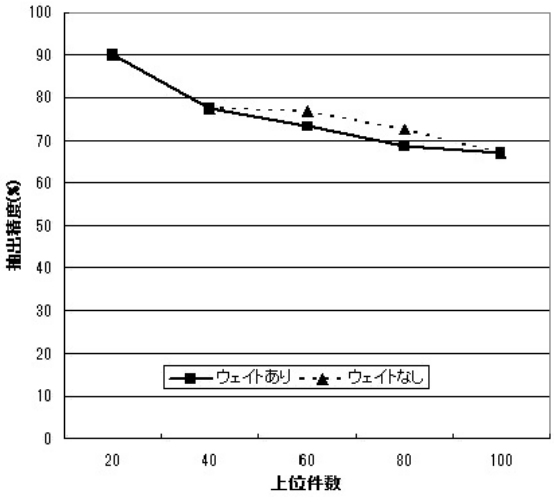

図 3: 抽出精度（検索要求者数ウェイト有無の比較）

語中で約 51 %〜67 %の語が正しく抽出された．この精度はいずれの年においても抽出数の増加と共に概ね逓減している．誤抽出されるものの多くには二項クエリ構造で第 2 項が省略され第 1 項が 2 語となるクエリが見られた（例：「名古屋 レストラン」など）．したがって第 1 項の構造を判別することが今

後の課題となる．

検索要求者数ウェイトの有無による比較では前者の方がより利用者ニーズの強さを反映した順位付けがされていると考えられる．しかし実験の結果，両者の差分は少ないことが確かめられた．上位 3，7，12 タイプ以内では両者に含まれる情報タイプの種類は完全に一致している（表 5）．また抽出精度においてもその差は最大 4 %以内に留まる（表 6）．

各情報タイプに属する情報には特有の検索処理や提示処理が必要なため，これらの検索・蓄積・提供に特化したサイトが既に存在することが予想される．実際に表 2 に含まれる情報タイプの大半に対しては有名なサイトが対応していることが分かる（表 7）．本提案手法により得られる主要な情報タイプのいずれかに対して専用のサイトが存在しない場合には，新たなサービスの開発余地があるといえよう．

表 7: 本手法で獲得した情報タイプに対する専門サイトの例

情報タイプ	対応する専門サイト例
動画	http://www.youtube.com（CGM）
レシピ	http://cookpad.com（CGM）
ブログ	http://www.google.com の Blogs 検索（検索サイト）
意味	http://dic.yahoo.co.jp/ （サービスサイト）
歌詞	http://www.uta-net.com/ （サービスサイト）
画像	http://www.google.com の Images 検索 （検索サイト）
wiki	http://ja.wikipedia.org/ （CGM）
価格	http://kakaku.com/ （サービスサイト）
ランキング	http://ranking.goo.ne.jp/ （CGM）
地図	http://maps.google.com （サービスサイト，CGM）

6. まとめ

本稿では，Web 上での専門サイトを特徴付ける情報タイプの重要性について論じ，これを検索クエリログから発見するための考え方としての二項クエリ構造モデルとこれを用いた情報タイプ抽出手法についても説明した．

また提案手法の有効性を確認するための抽出実験を行ったところ，上位 20 語で 90 %，上位 100 語で 67 %の精度であった（2010 年度データの場合）．精度を低下させる原因としてクエリ中で第 1 項が複数語をとり第 2 項が省略されたケースによるものが確認された．

情報タイプ辞書の構築に向けて，クエリの項構造を適切に解析し抽出精度を向上させることが今後の課題である．

謝辞

本研究は，文部科学省科学研究費基盤研究 (A)（22240007）の助成を受けて行われたものである．

参考文献

[Jansen98] Jansen, B.J., Spink, A., Bateman, J. and Saracevic, T. : Real Life Information Retrieval: A Study of User Queries on the Web, SIGIR Forum, Vol. 32, No. 1, pp. 5-17, 1998.

[Wen 02] Wen, J., Nie, J. and Zhang, H.: Query clustering using user logs, ACM Trans. Inf. Syst. (ACM TOIS) Vol. 20, No. 1, pp. 59-81, 2002.

[Arita 07] 有田一平, 菊池英明, 白井克彦：検索語の共起情報を利用した単語クラスタリングと Web 検索への応用, 情報処理学会研究報告, Vol. 2007, No. 76, pp. 115-120, 2007.

[Yasukawa 07] 安川美智子, 横尾英俊 : クエリログから獲得した関連語のクラスタリングに基づく Web 検索, 電子情報通信学会論文誌 D, Vol. J90-D, No. 2, pp.269–280, 2007.

[Nakawatase 11] 中渡瀬秀一, 大山敬三 : 検索クエリにおける修飾構造の調査, 電子情報通信学会技術研究報告, Vol. 110, No. 407, pp. 49-52, 2011.