

## 情報爆発時代の情報検索結果俯瞰サービスの提案

嶋津恵子  
慶應義塾大学  
shimazu@z6.keio.jp

森田想平  
慶應義塾大学  
ok.souhey@gmail.com

奥村祐介  
慶應義塾大学  
chai@sfc.keio.ac.jp

森薫  
慶應義塾大学  
kaoru.mori@gmail.com

**Abstract**

我々は、インターネット/イントラネットの検索エンジンが出力した結果を、鳥瞰的に俯瞰するデジタル・コンテンツ・アグリゲーションサービス(Digital Contents Aggregation Service)の実現を目指している。俯瞰用の metrics として、5W1H (Who, When, Where, What, Why, How) が有効であると考えており、今回、When 要素を使ったシステムを試作した。つまりコンテンツが持つ時情報を使って、検索結果コンテンツを年表上に俯瞰する機能の実現である。従来のランキングによる検索結果のリスト表示方式による特定のコンテンツ探索サービスと比べ、検索結果コンテンツ群の俯瞰サービスは、利用者になかなかな気づきを与える可能性が高まると期待される。

**1. Introduction**

Internet における情報流通に革命をおこした PageRank アルゴリズムは、今や利用者から満足されてはいない。これは、現在の(1)Rich-get-Richer 現象の発生と定着[2]、また(2)ブログ、CGM などの非専門家による解説や意見を適宜活用したいという要求への対応が検索エンジンに求められていること、さらに(3)非テキスト(音声、画像、動画)まで想定した検索結果に対する高度な活用の期待が理由である。これを受け、個々の情報に重要度を付与しリスト形式で出力する以外の方法が、模索され始めている。例えば、人工知能学会では情報編纂 (Information Compilation) という基盤技術の研究開発が提案され専門の研究会が発足した[3]。ここでは、情報の理解と効率的なアクセスを支援するシステムの構築に貢献する基盤技術の開発が目標となっている [4]。我々は、この情報編纂技術基盤を検索結果の俯瞰に応用することが狙いである。この時、情報編纂のこれまでの成果が、利用領域に特定した編纂手法の構築であるのに対し、本研究は、一般の検索エンジンの出力結果を取り扱うことを前提とする。そこで、検索結果を鳥瞰的に俯瞰する汎用的な metrics の開発を目指す。

検索結果をこの指標上に配置することで、①ロングテール部分に位置してしまう重要な情報を取り出しやすくすることと、②無秩序であった情報間の関係性の示唆を実現する。本研究は、利用者の知的な発見や気づきを誘発するという、知識創造に直接貢献する実用サービスの開発を目指す。

特に今回は、新表示方法の提案として metrics 候補中の時情報に注目した。年表を使い、検索結果コンテンツをそれらが持つ時情報で俯瞰するシステムの実現を試みた。これにより、利用者は検索キーワードに関連する出来事の変化、変遷、歴史を把握することができる。また、複数コンテンツの内容を横断的に表示することで、本来無関係なコンテンツ間の関係性を示唆することができ、知的創発に貢献できると考えた。我々は、実用性のある年表を生成するために、年表上に表示する効果的な「ヘッドライン」を、形態素解析結果から時に関係した重要語を抽出することで生成した。

**2. 年表表示状の課題**

情報編纂用の汎用的な metrics として知られているものに 5W1H (Who, When, Where, What, Why, How) があり、これを利用した報告が古くからある[5][6]。この中の Where を用い検索結果を鳥瞰的に俯瞰するサービスが、現在よく知られている GoogleMaps である。これは、コンテンツ中に出現する”場所を示す語”を使って地図上に該当するコンテンツの所在を用表示する。この時、”場所を示す語”は緯度経度表現に一意に変換可能である。このサービスは、一般に①緯度経度が特定できるような場所情報をコンテンツに記載するのは、明確な(顧客勧誘等の)目的がある場合に限られることと、②検索キーワードと任意の場所が、強い関係性を持つという2つの特徴を、有効に利用している。特に後者に関し、地図上に示された任意の場所と検索キーワードとの関係を、利用者は直感的に把握しやすい。図1の例では、検索キーワード「東京タワー」を入力したことで、実物の所

在地と関係施設(と推定される建築物)の所在地を把握することができる。

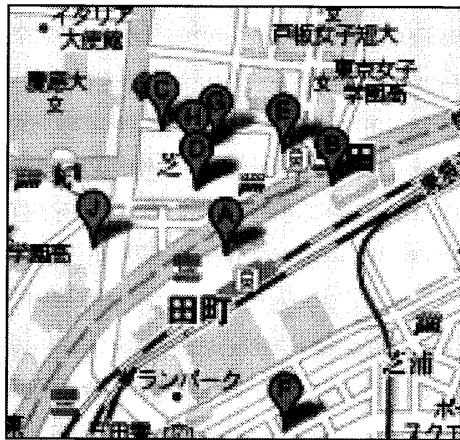


図1「東京タワー」の検索結果を地図上に配置

これに対し、特定の“時”と検索キーワードに強い関係性があることは(“場所”に比べ)少ない。例えば、図2は、前述の例と同じ「東京タワー」で検索した結果を、検索結果コンテンツが持つ時情報を使って、年表上に俯瞰したものである。地図上に配置したものと異なり、検索結果を有効に利用させる助けになっていない。

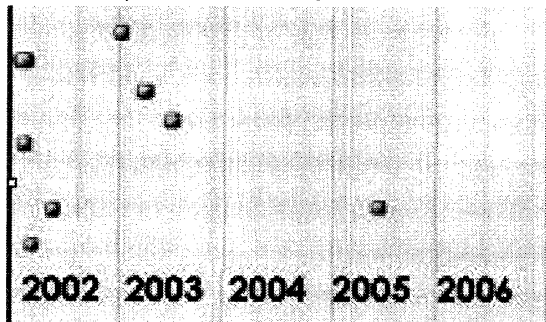


図2「東京タワー」の検索結果を年表上に配置

そこで我々は、検索キーワードとコンテンツに出現する時情報の関係を示す文言(ヘッドライン)を年表上に出力させる方法を検討した。

### 3. ヘッドライン生成と年表上への配置

我々は検索キーワードとコンテンツに出現する時情報の関係を示す文言(ヘッドライン)を、構文解析と固有名詞抽出技術を応用して生成した。具体的には、コンテンツ内の文章中の時情報を特定し、それと係り受け関係にあるその他の固有名詞と、同じく係り受け関係にある動詞と、さらに程度副詞を抽出し、それらの組を生成した。この語の組をヘッドラインとして採用した。

生成されたヘッドラインは、元文章が含まれるコンテンツのメタデータとして、検索エンジンが持つインデックステーブルに記録した。これにより、利用者が任意の検索キーワードを用いて検索を行うごとに、動的に年表上に表示することが可能になった。図3は、今回採用した方法でコンテンツを年表上に俯瞰出力した例である。

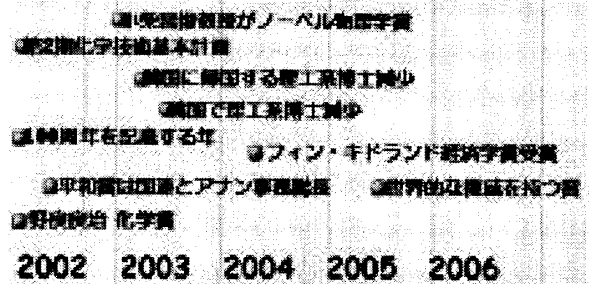


図3 検索結果をヘッドラインとともに年表上に配置  
検索キーワードは「ノーベル賞」

### 4. まとめ

我々は、汎用のアーキテクチャを採用した検索エンジンが出力した結果を、鳥瞰的に俯瞰するデジタル・コンテンツ・アグリゲーションサービス(Digital Contents Aggregation Service)の実現を目指している。俯瞰には、5W1H(Who, When, Where, What, Why, How)が有効であると考えており、今回、コンテンツが持つ時情報を使って、検索結果コンテンツを年表上に俯瞰するサービスを試作した。

### 8. 参考文献

- [1] Hawking, D. ; Web search engines . Part 1, Computer, Volume 39, Issue 6, pp. 86-88, 2006.
- [2] J. Cho and S. Roy, Impact of Search Engine on Page Popularity, Proc. of WWW2004, pp. 20-29, 2004.
- [3] 加藤恒昭, 松下光範, 情報編纂(Information Compilation)の基盤技術. 2006年度人工知能学会全国大会(第20回)論文集, 1D3-02, 2006.
- [4] 加藤恒昭, マルチメディアプレゼンテーションの自動生成に向けてー自然言語生成からマルチメディア生成へー, 情報処理, Vol. 38, No12, pp. 1049-1056, 1997.
- [5] M. T. Hansen, N. Nohria, and T. Tierney. What's your strategy for managing knowledge? In Harvard Business Review, volume 77(2), pages 106-116. 1999.
- [6] Shimazu, K. ; Arisawa, T. ; Saito, I. , "Interdisciplinary Contents Management Using 5W1H Interface for Metadata," IEEE/WIC/ACM International Conference on Web Intelligence,, pp909-912, 18-22 Dec. 2006