検索結果の年表表示に関する考察

A Verification of a search engine result mapping on a timeline chart

奥村 祐介† 嶋津恵子†
Yusuke Okumura† Keiko Shimazu†
†慶應義塾大学　デジタルメディア・コンテンツ統合研究機構
†Research Institute for Digital Media and Content, Keio University.

要旨
近年 web 上の情報量が爆発的に増大しているのに対し，大量の情報から必要な情報を探し出すサービスの開発が望まれている．本書では，インターネットの検索結果を，年表上に俯瞰させることの価値と，実現に向けての課題，およびその解決アプローチについて述べる．

1. はじめに

近年，インターネットの台頭とブロードバンド化の浸透が進んでいる．それにより，情報の発信やコミュニケーション活動がインターネットを使って行われる機会が増え，web 上の情報量は爆発的に増大した．

一方，大量な情報から必要なものを探し出す方法やそれを実現する情報技術に，大きな変化や進展はみられていない．インターネット上の検索エンジンは，Web 上の情報活用にパラダイムシフトをもたらしたが，今や万全の道具ではなくなっている．利用者がどんな情報を求めているかに関わらず，検索結果が画一的な重要度計算を用いて出力される．最適な情報を得るためには，利用者の“検索キーワード”(以下キーワード)の選定能力が重要となるが，複数の適切なキーワードを組み合わせたり，検索オプションを駆使することは彼らにとって難しい．また，リスト表示では，検索結果の下位(2 ページ目以降)に出力されるコンテンツは，ほとんど活用されない．米 iProspect 社によると，検索結果の最初の 3 ページしか閲覧しない利用者は 90%を越える[1]．インターネットの検索結果の情報を，より多く利用者に活用させるための主出力・表示方法が求められている．

本書は，インターネットの検索結果を，年表上に俯瞰させることの価値と，実現に向けての課題，およびその解決アプローチについて述べる．また，現在までに進展した検証を解説する．

本書は次の構成をとる．まず２章では関連事例と，時情報を用いた検索結果俯瞰機能の価値を整理する．３章に実現するための課題を示し、問題解決のアプローチを 4 章に述べる．5 章では今後の進捗について述べ，6 章にまとめを述べる．

2. 関連事例と検索結果俯瞰機能の価値

検索結果の俯瞰サービス事例として，既に多くの人に利用されている Google Maps[2] がある．Google Maps は，コンテンツに記載された場所情報に注目し，地図上にそれらを配置する(図 1)．

また，本書の先行事例として，Google の Alternate views for search result: Timeline View[3] (以下 Timeline View)がある(図 2)．このサービスは，検索結果のそれぞれに記載された時情報を使って，それらを年代毎に分類し，グラフ化をするもので、我々の課題に対する先行的事例である．[3]は，任意の年代を選択すると，それらに分類されたコンテンツに絞り込んだ検索結果がリスト表示される．このサービスは，キーワードが web 上でどれだけ取り上げられたか，その変化を把握することも可能である．一方，絞り込んだ検索結果がこれまでと同じリスト状に出力されるため，検索結果の一覧性が十分とは言えず，検索結果全体を俯瞰することには不向きである．

我々は，検索結果を一覧できるサービスを，年表上に検索結果を配置する方法で実現することを目指している．これにより，利用者は，任意の専門用語やイベント名の時系列の変化・変遷・推移を把握する助けになるに違いない．検索サービスは，大量なデータを対象に特定の情報を探し出すツールから，知識創発の武器になる．また，従来の検索結果の最大の利便性である任意の出力結果からコンテンツ実体にアクセスできる機能を踏襲することにより，年表上に検索結果を出力するサービスの効用は、さらに向上すると考える．

図 1. Google Maps
「東京タワー」で検索した出力結果

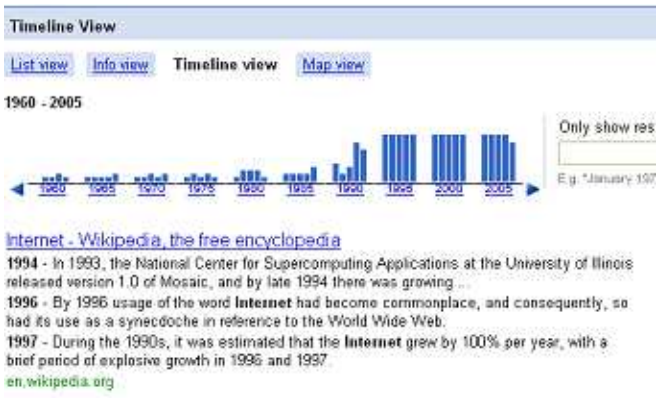

図 2. Alternate views for search result: Timeline View
「Internet」で検索した出力結果

3. 検索結果の時情報俯瞰実現に向けての課題

前章に挙げた有用性を実現するシステムを実現するためには，①年表表示に利用可能な時情報の特定と，②効果的な“ヘッドライン”の生成の，２つの手法の開発が必要だと考えた．それぞれに関し，以下に述べる．

3.1. 年表表示用の最適な時情報の特定

検索結果を年表上で俯瞰するためには，非構造である Web コンテンツから時情報の抽出が必要である．具体的には，非構造の文書からの時情報の特定と，それらから年表上に配置する際の目的に合致ものだけを選択するアルゴリズムの開発である．

一般に，コンテンツに場所(地図に配置可能な住所等)が記載される場合は，店舗の宣伝等の明確な目的があり，混乱を生むほど大量に抽出されることは無い．一方，時情報は，(コンテンツに表記しても(多くの場合)個人情報の開示になる可能性は低く)，多用される．このため，検索キーワードに内容的に無関係な時情報も抽出されることが懸念される．データベースを横断した検索結果を使った発見活動支援を狙いとした時，検索の目的に合致した時情報だけを正確に抽出するアルゴリズムの導入は，検索結果を年表上に表示するサービスの効果を最大化すると考えられる．

3.2. 時情報とキーワードの関連を示すヘッドラインの生成

検索結果のコンテンツから 3．1 節のルールで抽出した時情報を，年表上に配置した例を図３に示す．図 1(Google Maps)と比べ印を付けただけでは，有用性が低い．地図を用いた場合，利用者は検索に用いたキーワードとその場所の関係を直感的に理解できる．例えば図１では，（おそらく）東京タワーの位置が D であり，G，F，H はその関連施設だということが分かる．一方，年表上の固有の位置に印が付いた場合，キーワードとその「時」を示す文言がコンテンツに含まれていたことを理解するだけであろう．これでは，我々が前章に挙げた有用性が実現されているとは言えない．

これの解決策として，年表上の印と共に①コンテンツの Title タグの内容を表示する，②コンテンツの URL を表示する手法が挙げられる．(図４，図５)①は，一般的にタイトルはコンテンツの内容全体の要約や，無意味なものが多く、任意の「時」情報に有用な情報ではない．また，②も情報の発信元を推測する手掛かりにはなりうるが，「時」情報に有用な情報とは言えない．つまり，これらの方策は検索結果を年表上に配置したときの有用性向上に寄与するものではない．

そこで我々は，年表上で有用性を実現するためにはキーワードとその時情報の関係を示す用語を用いることを考えた（図 6）．時情報の位置にその情報を添えることで，利用者はキーワードに関連する出来事の変化や時系列上の分布を俯瞰することができ，利用者の新たな発見に繋がると考えた．

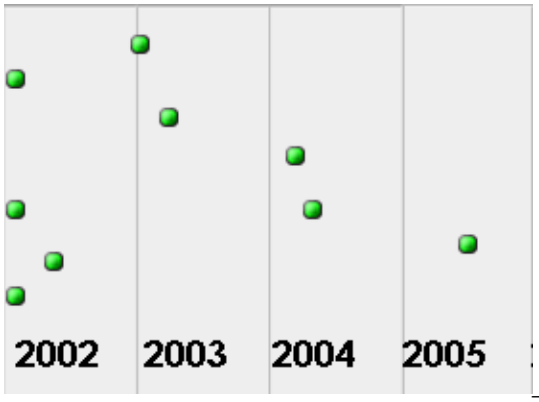

図３．「ノーベル賞」での検索結果から抽出した時 5 件のコンテンツを年表上に配置

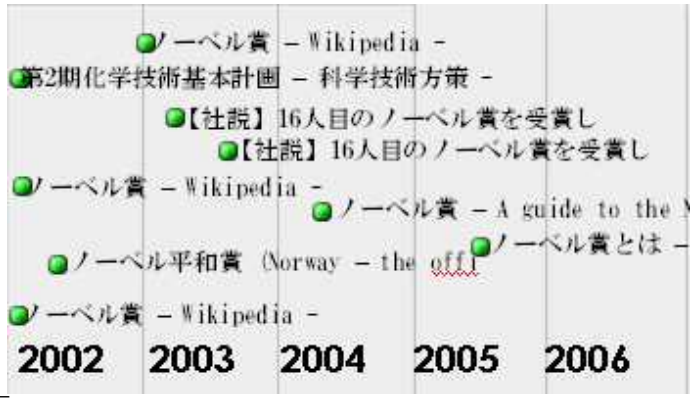

図４．抽出した時情報にコンテンツの Title タグ情報を付与

図５．抽出した時情報にコンテンツの URL 情報を付与

６．抽出した時情報に時情報とキーワードの関係を示す用語を付与

4. 実現へのアプローチ案

4.1. 時情報抽出：パターンマッチ

年表表示に利用可能な時情報の抽出と特定には，一般的なパターンマッチ手法を用いることを検討している．時情報を示す表現系として，年，月，日，季節，季語，時刻などが考えられるが，年表に表現することを前提とすると，年，月，日の組が最適であると考えられる．年月日が連続で出現するとき，これを年表に配置する時情報として抽出する．またこれらのうち 1 つもしくは 2 つが欠落する時は，それぞれの状況に応じて補足ルールを起動し，年，月，日からなる情報を生成する．具体的には次のとおり．

１．コンテンツに出現した時情報をそのまま利用

YYYY 年 MM 月 DD 日：YYYY/MM/DD に変換し，年表出力用時情報として採用

YYYY/MM/DD：年表出力用時情報として採用

YYYY-MM-DD：YYYY/MM/DD に変換し，年表出力用時情報として採用

YYYY.MM.DD：YYYY/MM/DD に変換し，年表出力用時情報として採用

２．コンテンツに出現した情報に特定情報を補完して利用

YYYY 年 MM 月：YYYY./MM/01 に変換し，年表出力用時情報として採用

YYYY/MM：YYYY/MM/01 に変換し，年表出力用時情報として採用

YYYY 年：YYYY/01/01 に変換し，年表出力用時情報として採用

３．コンテンツに出現した情報にサーバの発信情報を補完して利用

MM 月 DD 日：yyyy/MM/DD (ただし yyyy はサーバが示す発信年)に変換し，年表出力用時情報として採用

MM.DD：yyyy/MM/.DD (ただし yyyy はサーバが示す発信年)に変換し，年表出力用時情報として採用

MM 月：yyyy/MM/01 (ただし yyyy はサーバが示す発信年)に変換し，年表出力用時情報として採用

４．上記１，２，３，で生成した時情報のうち，(補完前の)元になる文言がキーワードを基準として前後N文字以内に出現した場合のみ，採用する；これはキーワードと関係の低い時情報の量産を避けることが目的.

4.2. ヘッドラインの生成：時情報を含んだ文の要約

ヘッドラインは，コンテンツ中に含まれる時情報とキーワードの関係を示す情報であり，検索結果を年表表示することの有用性を最大化する要素になる(3.2 節)．そこで、できるだけ性格にこの関係を示すことが求められる．その一方で、年表という限られた空間に文字列を配置することになるのでできるだけ短い字数で構成させるという制約も満たさなければならない．これを実現する方策として，自然言語処理(構文解析)結果の利用と，係り受け情報の応用を検討している．具体的には，次のとおり.

１．時情報と同じ係り先を持つ文節を結合

時情報が用言に係っている場合，時情報と同じ係り先を持つ文節を選択し繋げる（図７）.

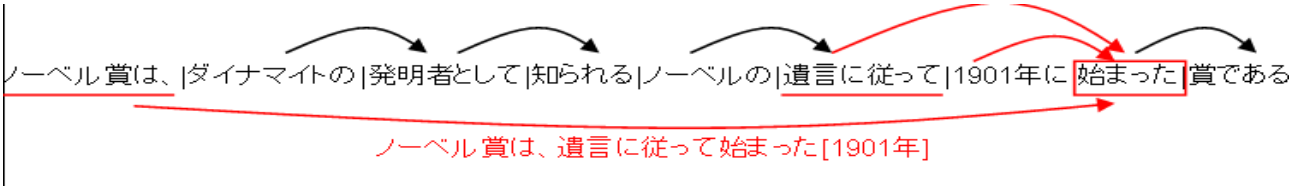

図７．時情報と同じ係り先(「始まった」)を持つ文節を繋げた要約例。(赤字のヘッドラインを生成)

２．時情報から係り先を辿り，初めに現れた用言を係り先に持つ文節を結合

時情報が用言に係っていない場合，時情報から係り先を辿っていき，初めに現れた用言を係り先に持つ文節を繋げる

.

5. 今後の方針

5.1. キーワードに関連する時情報の抽出手法に関する検証

本手法が，コンテンツ中からどれだけ正確に，もれなく時情報を抽出できているかを検証する．抽出した時情報と，手作業で作成した正解の時情報の比較を行い，精度再現率を測定する.

5.2. ヘッドライン生成手法の検証

要件に設定した，時情報とキーワードの関係が分かり，短く直感的に理解できるという点がどれだけ実現できているか検証する．要約文を年表上に表示し，有用性についてアンケート調査を行う.

6. まとめ

本書は，インターネットの検索結果を，年表上に俯瞰させることの価値と，実現に向けての課題，およびその解決アプローチについて述べた．また，その検証実験の概要仕様を示した.

参考文献

[1] iProspect Search Engine User Behavior Study.(April 2006)
http://*www.**iprospect**.com/premiumPDFs/WhitePaper_2006_**SearchEngineUserBehavior**.pdf*

[2] Google Maps　http://maps.google.co.jp/

[3] Alternate views for search result: Timeline View
http://www.google.com/views?q=view%3Atimeline