年表GUIの最大効果を目指した検索結果のフィルタリング

An Experimental Study of Filtering Research-Result for a Chronological GUI

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・コンテンツ統合機構
Research Institute for Digital Media and Content, Keio Univ.

要旨

GoogleMaps の成功を受けて，検索結果を通常のリスト形式以外の GUI 上に配置する手法が多く試みられている．特に情報共有の視点に立つと，「場所」を示す情報と同様に「時」を示す情報は重要だと報告されている． 一方，検索結果のすべてを対象にして，これらの GUI 上に俯瞰すると利用者のコンテキストに合致しないコンテンツも多く含まれることから効果が半減する．我々は，年表 GUI に検索結果を配置したときに，その効果がより大きくなることを目指し，検索結果のフィルタリングを実装した．さらにそれをイントラネット用検索エンジンに統合し，ニュース記事を検索したときに，年表 GUI 上への検索結果コンテンツの配置を実験した．

1. はじめに

Internet 上の検索サービス www.google.com（日本では主に www.google.co.jp）が検索サイト最大手となって久しい[1]．同サイトでは，Web ドキュメントの単純検索だけでなく，目的別の検索サービスも提供されている．その代表が GoogleMaps と呼ばれる地図検索サービスである．検索結果が地図上に表示される他，(Web 検索とは別のサービスとして)世界中の地図と衛生写真からなるコンテンツの参照サービスを提供している．検索結果の地図上への表示サービスは，指定した地域の店舗等をそれらの住所をキーとして，電話番号や住所とともに地図上に配置する．道順検索に利用され利便性が高いことから，利用者が急増し，今年に入ると利用者が作成した地図を検索する機能や，特定のコンテンツ群に限定し開催都市をキーとして検索するサービスを追加した[1][2]．一方情報の共有や活用の議論において,しばしば 5W1H(what, who, when, where, why, how)の重要性が取り上げられる[3][2]．我々はコンテンツに記載されている“時情報”(when)に注目した．営業活動の日報情報やコンテンツなどを俯瞰する際の日時情報の重要性は認識されてきた[4]．我々は，このように特定のデータベースのコンテンツに対し必要性が認められている“時情報”や，これをキーとして並び替えや検索する機能が，（任意のコンテキストに従って）検索された結果のコンテンツ群に対しても効果を発揮すると考えた．例えば，アルカイダに関する世界中のニュース記事を検索し，これらを発信日時別に並べなおすことで，活動の傾向を把握できる．

一方，実装アルゴリズムに関しては，検索結果コンテンツ群をそれらが持つ位置情報をキーに，地図上に配置する機能それは，比較的単純に実現される．具体的には，コンテンツ中に住所として認識できる表現があった場合，もしくはそれを何らかの方法で検索できるランドマークの表現があった場合，その住所から経度緯度を計算し，地図上に印を表示する．このとき，インタネット上の(誰もが参照できる)コンテンツに“住所[3]”を記載することそのものは，個人情報や守秘義務情報でもあり一般的でない．換

[1] アジアは若干傾向が異なり、特に日本では www.yahoo.co.jp、中国では www.baidu.com が最大手。
[2] 特にマーケティング(市場開拓)やシステムエンジニアリングの領域では､5W1H+H や 5W1H3C が議論されている。前者の 1H は How Much であり、後者の 3C は cost, computer, company を指す。
[3] より正確には，法律関係を処理する場合の基準となる場所であり、一般通念上は郵便物が正確に到着する表示を指す。

言すると，住所が明記されている場合は，コンテンツの発信者が何らかの(営業活動やサービス提供等の)目的で，利用者にその住所に訪れてもらうことを意図している．従って，1 つのコンテンツや情報に対し，これらの目的以外の“住所”がさらに別に記載されていることはほとんど無い．ところが，“時情報”(when)は，日常的に取り交わす一般的なものの一つであり，単一のコンテンツや情報に，様々な目的や狙いで複数出現することは珍しくない[4]．そこで，同一コンテンツに出現する“時情報”のうちどれを対象にするかの決定処理を加えないと，コンテンツの表示順位変え等のサービスを実現しても無意味な結果が多くなり利便性が悪くなる[5]．

我々は，慶應義塾大学のネットワーク上に存在するコンテンツを対象に，ニュース記事であるコンテンツだけを検索し，さらにそれらが持つ発信日時である”時情報”を選別した．そしてそれら”時情報”を対象に年表上に配置するサービスを実現した．

本書は次の構成をとる．2 章に我々が実装したサービスの概観を示す．3 章では選別を行わなかったときとの比較検証実験とその結果を示し，4 章に，結果の考察を述べる．5 章まとめと今後の発展を示す．

2. ニュース記事の年表配置サービス概要

2.1. 検索エンジンの構成

インタネットおよびイントラネットの検索エンジンの多くは，図１に示すシステムアーキテクチャを採用している[5]．クローラ部で収集されたネットワーク情報のコンテンツの情報は，一旦インデクサ部のリポジトリに蓄えられた後，検索システム毎に定義するデータフォーマットに整形され，検索部のインデックスとして格納される．このインデックス内のデータが利用者による検索対象として用いられる．このとき検索効率向上の目的から，検索結果に表示される各コンテンツの概要文(スニペット)データベースと，検索結果コンテンツの表示順位を決定するアルゴリズム(静的なランキング情報)を別に管理することが多い．特に google は，PageRank という手法を開発し，この静的なランキング情報モジュールに利用したことにより，利用度の高いコンテンツ出力順位を実現したことで有名である．我々が実現したニュース記事検索結果コンテンツの年表配置サービスは，この標準構成を拡張したものである．

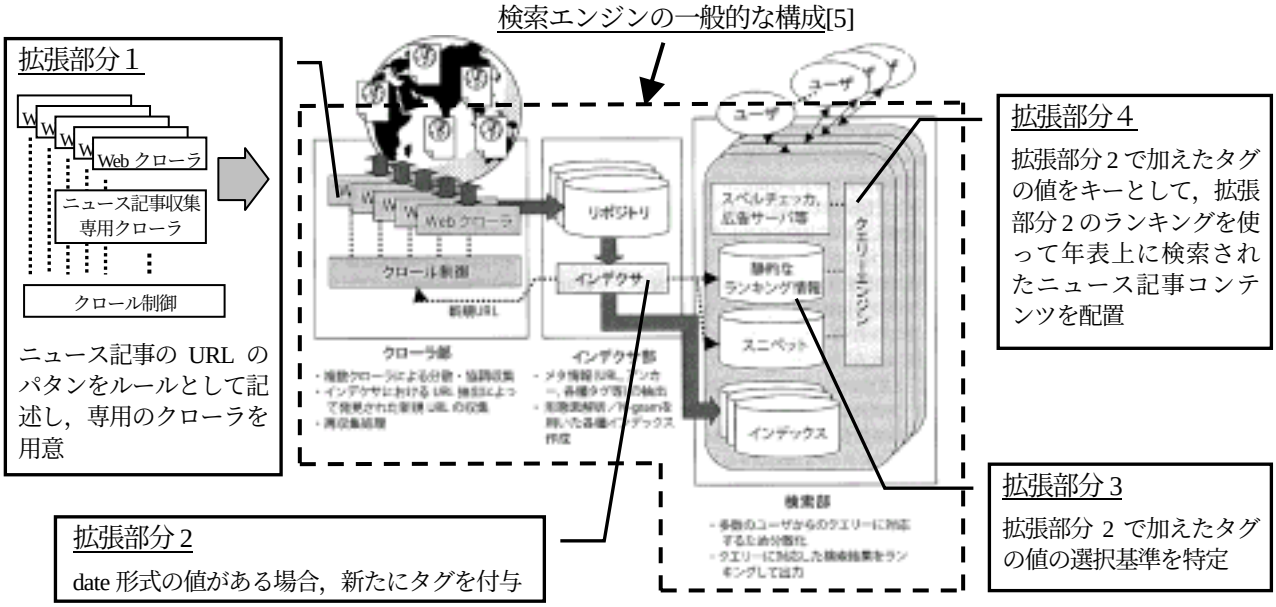

図 1　検索エンジンの一般的な構成 と 拡張部分

[4] 例えば，台風のニュース記事には，台風発生日時と日本上陸推定日時の２種類の時情報が掲載される．
[5] 例えば，上記の例では，台風発生日時か日本上陸推定日時のどちらかを選択し，コンテンツを並び換えないと利用目的に合致しない．

2.2. ニュース記事検索結果コンテンツの年表配置サービス概要

我々はニュース記事検索結果コンテンツの年表配置サービスを，検索エンジンに図 1 に示す 4 つの拡張を加えることで実現した．採用した検索エンジンの構成は，標準を採用した．

拡張部分 1：ニュース記事専用クローラの設置
この拡張部分は，クローラが収集する対象を，慶應ネットワーク上のすべてからニュース記事ページだけにふるい分けを行う．具体的には，事前にニュース記事として特定できる Web ページをサブネットごとに事前に 20 ドキュメントずつ特定し，それらに共通する URL パタンを人手で抽出した(図２)この後，専用のクローラを用意し，巡廻と情報収集規則にこのパタンを設定し，通常のクローラとともに，日常的に巡回運用させる．

拡張部分 2：検索対象データベースに専用タグを付与
IEEE の規格に準拠した日付型書式(mm/dd/yyyy や yyyy/mm/dd 等)の値がコンテンツ中に存在した場合，専用のタグ(“newtag_date”)を加える[6].

拡張部分 3：タグ newtag_date の値の選択順位の決定
任意の ID で特定されるドキュメント情報の中に newtag_date が複数存在する場合，次の優先順に基づいて一つだけを選抜する．(1)発信元 Web サーバがコンテンツ搭載/更新時に発行するメタデータ内の値，(2)ドキュメントを示す URL/URI 表現内の値，(3)コンテンツ本文内に最初に出現する値．選択されたものを並び替えの基準として用い，コンテンツが検索されたとき年表表示する．

拡張部分 4：検索されたニュース記事の年表上への表示
我々が提供する検索サービス上で，検索キーワードとして”ニュース”と別の 1 つ以上の語が入力された場合，拡張部分 1 のクローラが収集した結果のインデックスだけを対象に，“ ニュース ” を除く他の検索キーワードが含まれるコンテンツを検索する．そして拡張部分 2 および 3 の処理をした後，外部サービスである年表 GUI モジュールを起動し，年表上にコンテンツの要素情報を表示する．ここでの要素情報とは，年表上に表示する情報のことである．検索結果コンテンツの年表表示の仕様は次のとおり．
①検索されたコンテンツそれぞれの”タイトル”, ”拡張部分 3 で決定された日付情報”, ”概要”, ” URL” を検索用インデックスから抽出する． ②年表型ＧＵＩ起動 web サービス（今回は MIT が開発した SimileProject の Timeline ライブラリを利用）を起動[Timeline の参考文献]．③拡張部分 3 で決定した日付情報を基準として，検索結果コンテンツを年表上に配置する．
つまり、①で抽出した情報を XML 形式で Timeline ライブラリに与えることで，年表表示を実現する．

keio150.jp/news/
keio150.jp/events/schedule.html
www.keio.ac.jp/more_top.html
www.keio.ac.jp/more_campus.html
www.sfc.keio.ac.jp/visitors/news/
www.sfc.keio.ac.jp/students_soukan/news/
www.sfc.keio.ac.jp/alumni/news/
www.sfc.keio.ac.jp/faculty/news/

図 2．ニュース記事ページの URL パタン

3. コンテンツと時情報の選別の有効性検証

我々は，年表 GUI に検索結果を配置したときの効果がより大きくなることを目指した．そのために事前に，(a)ニュース記事だけをフィルタリングすることと，(b)どの “ 時情報 ” を利用するかの順位を指定することを提案している．そこで，これらを実施した場合としない場合の比較を試みた．具体的にはイントラネット用検索エンジンに 2 章で示した拡張部分の 1 から 4 すべてを統合した場合と，2 のみを利用場合で，検索結果に占めるゴミ情報(目的に合致しないコンテンツ)の割合率を比較した．

想定する利用ケースは, 任意のコンテキストのニュース記事を発信(もしくはニュース記事で報道されているイベントの実施)日時順に俯瞰するものである．従ってゴミ情報の判定は，(i)検索されたコンテンツの内容がニュース記事で無い，もしくは(ii)リストされた順位がニュースの発信日時では無いの，いずれか(もしくは両方)に合致するかどうかである．10 例の検索をおこなった．たとえば, “ ニュース ”“ 受賞 ”で検索すると，第一検索キーワード(“ニュース”)により，我々が実装した検索サービスは，通常の検索用インデックスではなく, ニュース記事専用のそれから第二検索キーワード(“受賞”)を含むものを抽出した．フィルタリングを行わない場合，下表に示すとおり，検索結果コンテンツのうちゴミ情報はおよそ 99%に達した.

表．ニュース記事の絞り込みを行わなかった時の不要なページ数

第二検索キーワード	ページ合計数	有効ページ数	不要なページ数
シンポジウム	99	6	93
会議	500	0	500
公告	500	0	500
受賞	500	8	492
特集	500	8	492
締結	500	8	492
講座	500	8	492
講演	500	8	492
講義	500	2	298
開催	500	8	492

4. 考察

実験の結果から，我々が提案するフィルタリングが有効であるといえる．その一方で，選別が 2 箇所で実施されていることに注意したい．1 つ目が検索用のコンテンツ収集であり，2 つ目がコンテンツに記載されている日付表現の選択である．具体的には，前者はあらかじめニュース記事だけを検索対象とするインデックスの作成であり，後者はそれらニュース記事を発信日時順に並べるための最適な日時情報の選択である．今回のゴミ情報の除去の多くは，前者によって実現されているものを思われる．我々の提案の本質は，目的に合致する最適なものを選ぶための日付情報の選択である．これは，GoogleMaps に代表される地図上への検索結果コンテンツの配置のときと異なり，日付情報は同じコンテンツ上に頻出することによっている．従って，我々の提案を精度高く評価するためには，日付情報の選択の処理部分を統合した場合とそうでない場合の比較をさらに実施する必要がある.

5. まとめと今後の計画

我々は, イントラネット上のコンテンツ検索結果を年表 GUI に配置することで地図上に配置するサービスと同様の効果を発揮することと，ただしそれは的確な情報の絞込みが必要であることを提案した．実際に検索結果のフィルタリング機能を実装し，それをイントラネット用検索エンジンに統合し，ニュース記事を検索のケースを想定し，実験をおこなった．この結果から我々の提案が有効であることを確認した．本発表以後，速やかに 4 章に示す追加実験を行う予定である.

参考文献

[1] http://www.itmedia.co.jp/bizid/articles/0710/11/news011.html
[2] http://www.itmedia.co.jp/news/articles/0710/03/news019.html
[3] 奥村 明俊, 池田 崇博, 村木 一至, MIIDAS : 情報の選別的共有のためのオントロジ構築とその増進的学習, (社)情報処理学会第 55 回全国大会講演論文集, pp. 240-241 (2000)
[4] http://marketing.mitsue.co.jp/archives/000123.html
[5] 山名 早人, 村田 剛志, 検索エンジンの概要, 情報処理, 特集 検索エンジン 2005 –Web の道しるべ-, Vol.46 No9, pp. 981-987 (2005)
[6] ISO 8601 Data elements and interchange formats -- Information interchange -- Representation of dates and times