修士論文 2008年度 (平成20年度)

検索結果コンテンツの年表型俯瞰システム

慶應義塾大学大学院 政策メディア・研究科

氏名：奥村 祐介

修士論文要旨 2008 年度 (平成 20 年度)

検索結果コンテンツの年表型俯瞰システム

論文要旨

本研究は，検索結果コンテンツを年表上に出力するサービスの実現を目的とする．これにより，検索結果コンテンツ群を年表上に編纂することで利用者に知的創発を促すことを目指している．

近年検索エンジンは，コンテンツの大容量化への適応と検索高速化の期待に応え，近年飛躍的な発展を遂げ，その基盤技術は成熟した．現状，代表的な検索サイト上では，世界中の Web コンテンツを対象に 1 秒未満で検索結果が出力されるまでになった．

一方，検索結果の表示方法に関しては大きな発展はみられていない．検索結果の表示手法は，コンテンツのサマリと URL からなる情報の組を表示する手法が主流となっている．しかし，利用者の目的に関わらず，検索結果が画一的な重要度計算を用いて出力されるため，最適な情報を得るには，利用者はキーワードを適切に選定する力を求められる．また従来の出力方法では検索結果の最初の 1 ページ (10 件) しか参照しないユーザは 60%，最初の 3 ページ (30 件) しか閲覧しない利用者は 90%を越えているという調査結果も発表されている．これは，爆発状態にある情報のほとんどにアクセスできていないことになる．つまり，自身の目的に照らすと重要であるコンテンツを探し出せない可能性はますます大きくなる．

そこで本研究では，新表示方法の提案として「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．本研究では，実用性のある年表を生成するために (1) 年表に出力する最適な時情報の抽出と (2) 効果的な「ヘッドライン」 生成の，2 つの手法を開発した．(1) の解決には，パターンマッチを用い，(2) 形態素解析結果から「時」に関係した重要語を抽出した．

それぞれに対し有用性を検証し，本手法による時情報抽出とヘッドライン生成手法の組み合わせは，検索結果コンテンツの年表を使った編纂に有効であると分かった．

キーワード
1. インターネット，2. 検索エンジン，3. 情報俯瞰, 4. 年表表示

慶應義塾大学大学院 政策・メディア研究科

奥村 祐介

Abstract of Master's Thesis

Academic Year 2008

A Chronologizing System for Search Results

Abstract

The goal of this research is to achieve a service which plots search result contents on a timeline. By compiling the search result contents into a timeline, we aim to stimulate the intellectual emergence by users. In the recent years, search engines have rapidly developed to meet the expectations of adaptation to large volumes of contents and speed of search, and the platform technology has matured. Currently, the major search websites return an output of web content search results within 1 second. However, there has not been much progress in the methodology of search result representation. It is currently common to represent the search results by displaying a pair of information which consists of the summary of contents and its URL. Regardless of the user 's objectives, search results are created by uniformly calculating the information. Therefore, to acquire the optimal information the users are required to select the appropriate keywords. In addition, there are surveys which report that 60% of the users only refer to the first page(10 articles) of search results, and over 90% only refer to the first 3 pages (30 articles) with the conventional method. This indicates that the users rarely have the access to most of the information which is growing at an explosive rate.

In other words, there is a higher possibility that the users are not able to find the contents which are important to their objectives. In this research we propose using a " timeline " to provide an overview of the information using " time ". This allows users to recognize the change, transition and history of events related to the keyword that the user provides. By representing multiple contents in a cross-cutting manner, we can also present the relationships between information which are normally unrelated, thus contributing to the intellectual emergence.

To create a practical timeline in this research, we have developed methods to 1) extract optimal information of time that would be presented on the timeline, and 2) create an effective " headline ". To solve the problem of 1) we have used pattern matching, and for 2) we have extracted important terminology concerning " time " from morphological analysis results. We have verified the effectiveness of each method and concluded that the combination of our time-information extraction method and headline creation method is effective for compilation of search result contents using timelines.

Keywords :

1. Internet, 2. Search engine, 3. Overviewing, 4. timeline

Keio University , Graduate School of Media and Governance

Yusuke OKUMURA

目 次

第 1 章　はじめに **1**
1.1 背景 1
1.2 本研究の目的 2
1.3 本論文の構成 3

第 2 章　関連プロダクト・関連サービス **4**
2.1 関連プロダクト 4
2.1.1 Simile Project: Timeline 4
2.1.2 today's timetable 5
2.2 関連サービス 5
2.2.1 Google Maps 5
2.2.2 Google TimelineView 6
2.2.3 ニュース年表 7
2.2.4 AllofMe 8
2.3 本章のまとめ 8

第 3 章　検索結果の年表俯瞰実現上の課題 **10**
3.1 年表に出力する最適な時情報の抽出 10
3.2 時情報と検索キーワードの関係を表す “ヘッドライン” の生成 11
3.3 本章のまとめ 12

第 4 章　提案手法 **14**
4.1 年表に配置する文の適正 14
4.2 時情報の抽出：パターンマッチ 15
4.3 ヘッドラインの生成：時に関係する重要語の組 16
4.3.1 重要語の抽出 16
4.3.2 ヘッドライン生成ルールと適用順序 18
4.4 本章のまとめ 20

第 5 章　検索結果年表表示システム概要 **22**
5.1 システム構成: 汎用検索エンジンの拡張 22
5.2 処理フロー概要: 検索エンジン・indexes を利用 22

第 6 章　有効性の効果検証 **25**
6.1　抽出した時情報の有用性 25
6.2　生成したヘッドラインの有用性 25
6.2.1　先行手法との比較 26

第 7 章　考察 **29**
7.1　時情報抽出について 29
7.2　ヘッドライン生成について 30

第 8 章　結論 **31**

図 目 次

1.1 検索エンジンの一般的な構成 [1] 2
1.2 本研究の目指す出力例 - ノーベル賞の検索結果コンテンツの内容を年表上に表示 3

2.1 Simile Timeline の出力例 - ケネディ大統領暗殺の日を解説する表 . . . 4
2.2 today's timeline の出力例 - 公演のタイムスケジュール 5
2.3 Google Maps の出力例 - "東京タワー"で検索した出力結果 6
2.4 Google TimelineView の出力例 - "Internet"で検索した出力結果 7
2.5 ニュース年表 の出力例 - 「オリンピック」で検索した出力結果 7
2.6 AllofMe の出力例 - 上は「アップル社製品の歴史」, 下は「マイクロソフト社製品の歴史」を表示 8

3.1 ノーベル賞の検索結果から抽出した 5 件のコンテンツを年表上に配置 . 11
3.2 左:抽出した時情報に Title タグ情報を付与　右:URL 情報を付与 12
3.3 時情報に時情報とキーワードの関係を示す用語を付与 12

4.1 年表上に表示した時に有用性の高い文の例 14
4.2 年表上に表示した時に有用性の低い文章の例 14
4.3 固有名詞の抽出 17
4.4 括弧内の文言の抽出 17
4.5 程度副詞とその前後の名詞の抽出 18
4.6 時の係り先の用言とその関連語を抽出 19
4.7 ヘッドライン生成プロセス 20

5.1 既存検索エンジンの拡張 - システム構成図 [1] 23
5.2 年表表示機能の情報処理フロー 24

6.1 適用ルールの内訳 27
6.2 提案手法と Chasen の抽出数比較 27
6.3 提案手法と Chasen の精度比較 28

表 目 次

6.1 時情報のパタン抽出:精度再現率 25
6.2 ヘッドライン生成:正解における網羅率 26
6.3 提案手法と Cabocha の精度再現率 28

第1章　はじめに

1.1　背景

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．それに伴い，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は増え，web 上の情報量は爆発的に増大した．インターネット利用者が，この大量の情報を有効に活用するためには，目的の情報を見つけアクセスするための適切な手段が必要である．そこで現在，インターネット利用者が Web 上に存在する大量の情報の中から目的のコンテンツを探し出す際に利用されているのが，検索エンジンである．

検索エンジンは，Web 上を巡回してコンテンツを保存し，利用者の検索リクエストに応じて検索結果を出力する．利用者は検索エンジンに対し，検索キーワードを含めた検索式を指定することで，求める情報へのアクセスが可能となっている．一般的な検索エンジンの構成を図 1.1に示す．この図は，Web や DB からデータを収集し保存，検索高速化のためにインデックス化を行い，利用者のリクエストに応じて検索結果を返す一連の流れを示している．この検索エンジンは，コンテンツの大容量化への適応と検索高速化の期待に応える形で近年飛躍的な発展を遂げ，その基盤技術は成熟したと言える．ここで言う基盤技術とは，検索エンジンの基本機能である，Web 上から大量の情報を収集する技術，収集した情報を保存する技術，保存されている情報から素早く検索を行う技術の 3 つである．図 1.1ではそれぞれ Gathering，Extracting，Indexing の 3 つに該当する．そのどれもが利用者にとって満足できるレベルに達しており，現状，代表的な検索サイト [2][3] では，世界中の Web コンテンツを対象に 1 秒未満で検索結果が出力されるまでになった．

一方，基盤技術の成熟に対し，利用者への情報提供手法に関しては大きな発展が確認されていない．図 1.1では，利用者への Presenting の部分がこれに当たる．検索結果の表示手法には，コンテンツのタイトルやアクセス (URL)，検索語が含まれる文書の一部 (スニペット) をリスト形式で表示するという従来の手法が定番として定着し，利用されている．しかし，そのリスト形式での表示手法は，万全の道具ではなくなってきている．従来の手法では，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ため，最適な情報を得るためには，利用者のキーワードの選定能力が重要となる．しかし，複数の適切なキーワードを選択し，それを組み合わせたり，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 [4]．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．

本研究では，新表示方法の提案として” 時”を用いた” 年表”に焦点を当て，検索結果の俯瞰を試みる．俯瞰とは，” 高い所から見下ろすこと．全体を上から見ること”を意味しており，検索結果コンテンツの横断的な俯瞰は，本来無関係なコンテンツ間の関係性を示唆することに繋がると考えられる．年表を利用することは，情報共有の際に利用される，一般的によく知られた 5W1H(When, Where, Who, What, Why, How) の When の効果を検索結果出力にも応用する試みである．5W1H は古くから情報の整理学や情報の本質を共有する際に利用されており，その報告は古くから存在している．[5][6]

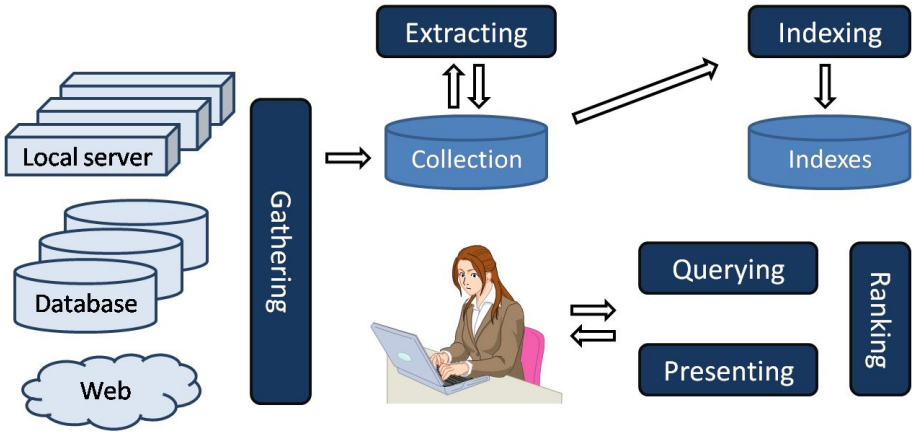

図 1.1: 検索エンジンの一般的な構成 [1]

1.2 本研究の目的

本研究の目的は，検索結果の内容を年表上に俯瞰させることで，利用者に知的創発を促すことである．実現を目指すシステムの出力例を図 1.2に示す．これは検索エンジンを用いて「ノーベル賞」というキーワードで検索結果コンテンツ群を取得し，その内容を元に年表を作成した例である．日本の受賞に関するニュースを赤の四角で囲み，ユニークなニュースを青の丸で囲った．これを見ると，例えば日本に注目した場合，受賞は物理学と化学のみであり，2002 年からしばらく受賞が無かったものの，2008 年に 4 人もの受賞者を出していることが一目で分かる．また，”2001 年がノーベル賞の 100 周年”であったり，” 漫画ノーベル賞”の存在，“日本がアジア地域で輩出数が 1 位”などの項目は，筆者にとっては新たな発見であった．これらはリスト表示であれば，検索結果コンテンツ順に読み，その項目に行き当たらなければ発見できないが，年表表示を用いた俯瞰では一目で概観できる．このように年表俯瞰は，利用者に任意の専門用語やイベント名の時系列の変化，変遷，推移を把握させる助けとなるに違いない．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，新たな発見，

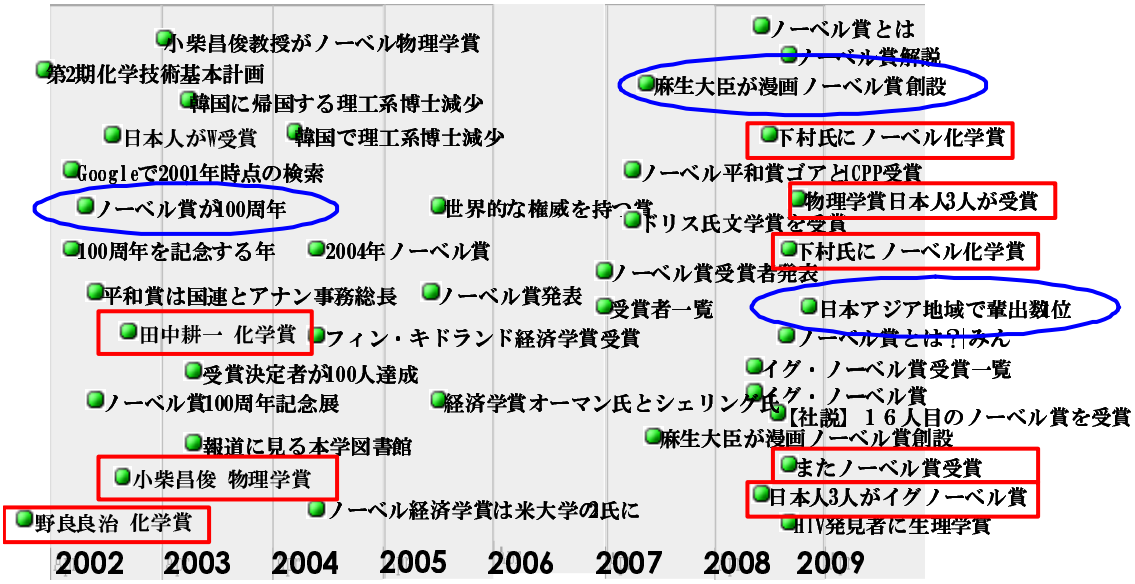

図 1.2: 本研究の目指す出力例 - ノーベル賞の検索結果コンテンツの内容を年表上に表示

知的創発の武器となる．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．

1.3 本論文の構成

本論文は 10 章で構成される。2 章では，本研究の関連プロダクト・サービスについて述べる．3章では，有用な検索結果コンテンツの年表表示のために必要な課題を特定する．4章では，3章で示した課題解決のための手法を提案する．5章では，年表俯瞰サービスを実現するシステムの構成について述べる．6章では，提案した手法の有効性と課題を導き出すため行った，実験検証結果を報告する．7章では，提案手法やシステム，実験結果を元に本研究を考察する．8章で本研究をまとめ，結論を述べる．

第2章　関連プロダクト・関連サービス

本章で，本研究の関連事例を挙げる．汎用のソフトウエアとして異なる目的に転用が可能なものをプロダクトとし，固有の利用目的を持ち特定のサイト上で提供されているものをサービスとしてまとめた．

2.1　関連プロダクト

2.1.1　Simile Project: Timeline

Simile Project の Timeline [7] は，MIT で開発されている，情報を時系列で表示するための Javascript ライブラリである．利用例を図 2.1に示す．この図は，ケネディ大統領暗殺の日の出来事を出力した例である．年表上にある一つ一つの点が，暗殺に関連する出来事を表しており，数分間隔での出来事の推移が見て取れる．年表下部は，上部に比べ年表の縮尺が大きくなっており，情報が圧縮されたかたちで閲覧できている．

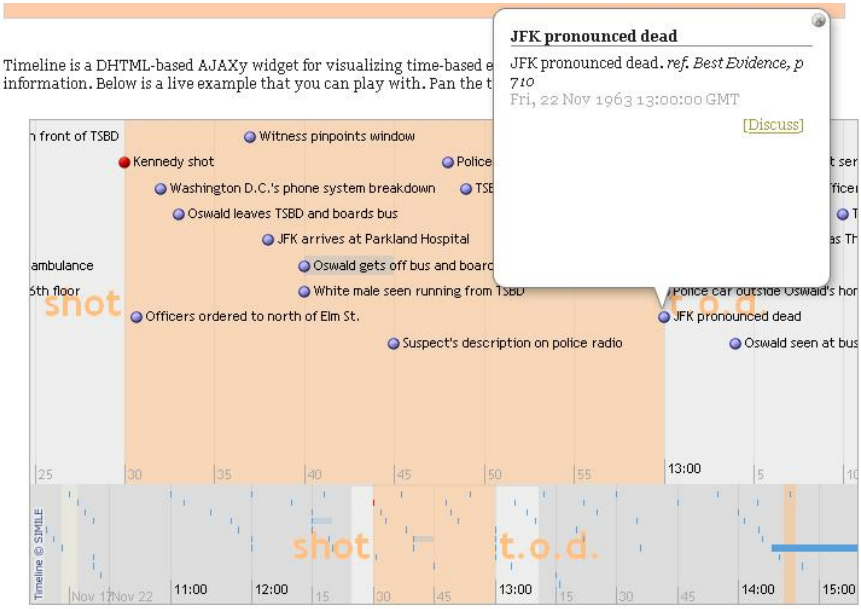

図 2.1: Simile Timeline の出力例 - ケネディ大統領暗殺の日を解説する表

利用者は，年表を自由にスライドさせることで，時間を滑らかに遷移しながらイベントの閲覧ができる．また，図 2.1のようにポップアップ機能を有しており，利用者は気になるイベントを発見した際に，より詳細な情報を即座に確認できる．ポップアップ部分は URL のリンクも含めることができるため，他コンテンツへのアクセスも容易である．

Timeline は，年表のサイズや縮尺，文字の大きさ，配色，横置き/縦置きなど多くのパラメータが調節可能であり，年表表示に対するライブラリ利用者の要求に柔軟に対応できる．また Timeline は，年表上に文字列を列挙していく際の効率的な配置を実現している．年表の空間に無秩序に文字列を配置してしまうと，文字が重なってしまったり偏った配置になってしまう問題が発生し，閲覧者にとっての一覧性が損なわれてしまう可能性が高い．Timeline は，入力されたデータに応じて，効率的にバランスよく文字列を配置する．

2.1.2 today's timetable

today's timetable[8] は，タイムテーブルを実装するための Javascript ライブラリである．利用例を図 2.2に示す．これは，公演のタイムスケジュールを，出演者の顔写真と共に表示した例である．さらに，各出演者の Wikipedia[9] ページへのリンクを表示させることで，プロフィールの確認が行えるようになっている．

図 2.2: today's timeline の出力例 - 公演のタイムスケジュール

today's timetable は，時刻表上に情報を表示させることができる．時刻表という点から，イベントのスケジュール表やテレビ番組表といったものを作るときに有用であると予想できる．リンクを埋め込んだり，マウスクリックでイベントを発生させることができるため，ライブラリ利用者は時刻表をスタートとして，様々な形での情報提供を検討することができる．

2.2 関連サービス

2.2.1 Google Maps

Google Maps[10] は，検索キーワードとそれに関連する位置を，地図上に示すサービスである．サービスの出力例を図 2.3に示す．これは，" 東京タワー" をキーワードに検

索した時の出力結果である．地図には東京タワー関連の場所に複数のピンが立っており，その位置が確認できる．

図 2.3: Google Maps の出力例 - “東京タワー”で検索した出力結果

Google Maps は，検索エンジンが出力する検索結果を，位置情報を利用した俯瞰サービスとして実現している本研究の先行的事例である．検索エンジンを用いて収集したコンテンツから，緯度経度を特定できるような住所情報を抽出し，地図上のその位置に印を打つことで，検索キーワードとそれに関連する位置を示す．

2.2.2 Google TimelineView

Google TimelineView[11](以下 TimelineView) 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．サービスの出力例を図 2.4に示す．これは，“Internet” をキーワードに検索した時の出力結果である．中央のチャートは，キーワードを含んだ検索結果コンテンツの中に，その年代を示す時情報がいくつあるか示したものである．例の場合，数値は 1985 年まで低いが，1990 年から増え始め 1995 年以降は高い数値を保っている．これは，Internet の検索結果には 1995 年以降に発生した出来事の記述が多いということを示している．さらに，チャートを元に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．

TimelineView は，絞り込んだ検索結果がこれまでと同じリスト形式で出力されるため，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と言える．

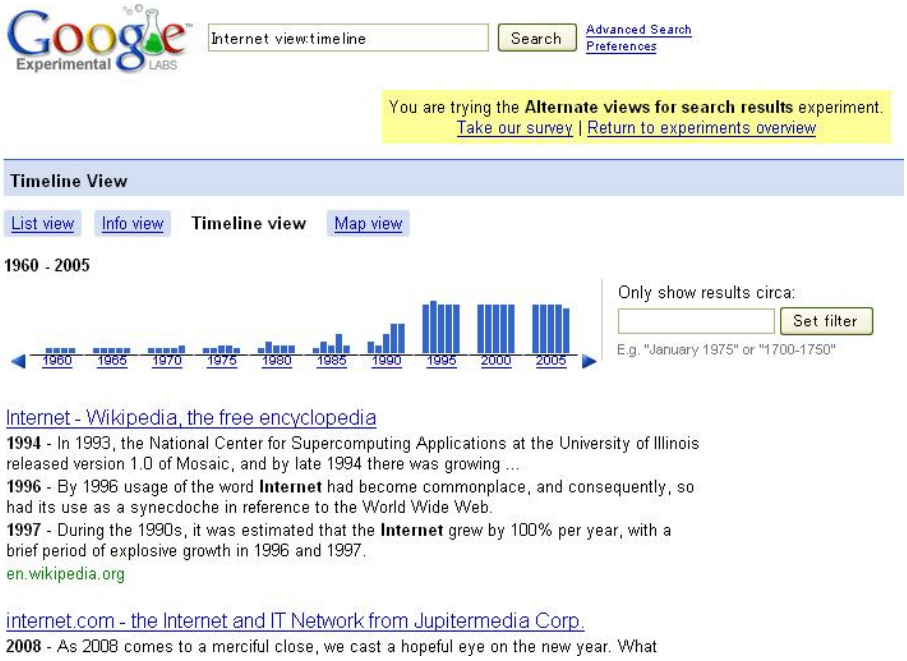

図 2.4: Google TimelineView の出力例 - “Internet”で検索した出力結果

2.2.3 ニュース年表

mediajam の運営するニュース年表 [12] は，ニュースサイトを対象にした検索結果を，年表を使って俯瞰させるサービスである．サービスの出力例を図 2.5に示す．これは，「オリンピック」をキーワードに検索した時の出力結果である．年表上には，ニュース発信日の位置にニュースタイトルが表示され，並んでいる．

図 2.5: ニュース年表 の出力例 - 「オリンピック」で検索した出力結果

ニュース年表は，検索対象をニュースサイトに限定し，検索時リアルタイムに発信日とタイトル，記事の URL，スニペットを抽出する．それを元に Simile Timeline[7] を用いて年表を作成し，表示している．

2.2.4 AllofMe

AllofMe[13] は，自分史を年表で俯瞰するためのサービスである．このサービスの有用な点は，自分史を作成し眺められるだけではなく，様々な人が登録した自分史や歴史を並べて見比べることができる点である．サービスの出力例を図 2.6に示す．これは上の年表に「アップル社製品の歴史」， 下の年表に「マイクロソフト社製品の歴史」を表示している例である．利用者は，異なる対象の歴史年表を並べ比較することで，対象の関連性を把握でき，これは新たな発見に繋がる可能性がある．また物事の関連性を利用して，音楽チャートの歴史などを表示し，記憶を辿りながら自分史を作成する，といった利用法も考えられる．

図 2.6: AllofMe の出力例 - 上は「アップル社製品の歴史」, 下は「マイクロソフト社製品の歴史」を表示

2.3 本章のまとめ

本章では，年表表示や検索結果の俯瞰に関連するプロダクト，サービスをまとめた．これらの事例は全てここ数年に登場したものばかりであり，利用者に対する新たな情報提供方法への期待が高まっていることが伺える．

年表表示に関するサービスを見てみると,「自分史」や「歴史」を対象とするものが多い．これらは年表にする価値が分かりやすく，年表表示の有用性が発揮できるコン

テンツ群である．しかし，ニュース年表を除いて，年表表示サービスはあらかじめ誰かが情報を登録し，それを閲覧するサービスであり，動的に年表を生成する試みはまだ始まったばかりだということが分かる．

Google TimelineView は，時情報を用いて検索結果を俯瞰させるという点が本研究と共通している．TimelineView は，Web 上で多く取り上げられた年代を示し指定させることで，ユーザに情報を絞り込ませ，目的のコンテンツへの到達を容易にすることが目的である．本研究は，情報を一覧性の高い年表上を用いてユーザへ情報を表示することで，別コンテンツ間の情報の関連性を示し，知識の創発を促すことを目的としている．この目的の違いが，リスト表示と年表表示という出力形式の違いに現れている．

ニュース年表は，ニュースサイトを対象としており，検索キーワードを含むニュースコンテンツを年表上に配置する．これは，検索結果を元に動的に年表を生成する先行事例である．ニュースサイトに対象を限定することには，以下の二つの利点があると考えられる．一点目は，誤った情報が混じる可能性が極めて低い点である．ニュースサイトの情報は，記者や編集者が作成しているものであり，情報の信頼性は一般のインターネット利用者に比べて高いと言える．ニュースサイトは，マスメディアとして迅速かつ正確な情報を伝える性格が強い．二点目は，ニュースコンテンツは一つの記事に一つの話題しかないため，その発信日とタイトルを抽出し表示するだけで，任意の時に何が起きたかを簡潔に示すことができる点である．さらに，ニュースサイトを限定し，サイト構造が把握できていればこの 2 つは比較的容易に抽出できるため，実現の難易度も高くない．本研究がニュース年表と異なる点は，対象とするコンテンツを制限しない点である．そのためニュース年表と比べると，信頼性の低い情報が混じる可能性が高くなり，予測できない構造のコンテンツから有用な年表を作成しなければならないため，実現難易度が高くなってしまう．しかし，対象を制限しないことで，ニュースサイト以外の情報源を活用した年表を作成できる．現在でも Web 上には，Wikipedia[9] や Wiki，知識人の blog や掲示板，有志が任意の情報をまとめたサイトなど，十分に有用な情報を持つコンテンツが多く存在すると考えている．有用な情報は今後も新しい形で発信される可能性があり，これら未知のコンテンツ群から，文章で表現されているコンテンツ全てを対象に年表表示を目指す本研究には重要な意味があると考えている．

第3章　検索結果の年表俯瞰実現上の課題

第 1.2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抽出と，(2) 効果的な “ヘッドライン” 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　年表に出力する最適な時情報の抽出

本研究では，検索結果コンテンツの内容を年表上に配置する．そのためにはまず，コンテンツ中から年表上の位置を示す時情報を特定しなければならない．具体的には，時の流れの一点を指す日付情報の抽出が必要となるが，これは困難な作業である．

困難である理由の 1 点目は，Web 上のコンテンツは時情報の形式やサイト構造が配信者によって自由に決定されているため，時情報の場所と形式が不規則である点である．出現場所に関しては，コンテンツ全体を検索対象にすることで問題は解決するが，形式に関しては，例えば 2008 年 11 月 1 日と伝えるだけでも 2008/11/1 や 2008-11-1 など，人によって様々な形式での記述が予想される．それらの時情報をできる限り正確に漏れなく抽出する技術が，必要となる．

理由の 2 点目は，時情報は一般的に出現回数が莫大であるという点である．Google Maps を例に比較した場合，一般的にはコンテンツに場所 (地図に配置可能な住所など) が記載される場合は，店舗の宣伝など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ない．一方、時情報はコンテンツに併記しても個人情報の開示になる可能性は低く，日常的に多用される言葉である．複数の検索キーワードの上位 100 件を対象に数の調査をしたところ，時情報は大抵数千を数えるが住所情報は数カ所から十数カ所であった．十数カ所であればすべてを出力しても大きな混乱は無いが，数千となると人間が直感的に理解できる数を超えてしまう．しかし，それらの時情報は，全てが利用者に有用な出来事を説明している訳ではなく，サイト更新日時やただ記事の発信日を示しているなど，表示しても無意味なものも含まれている．

検索結果を横断し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を実現する上で必要不可欠であり，年表の効果を最大化するために必要である．

3.2 時情報と検索キーワードの関係を表す“ヘッドライン”の生成

第3.1節で示したように検索キーワードに関係の深い時情報のみをコンテンツから抽出できたと仮定し，その時情報のみを，年表上に配置した例を図3.1に示す．これはノーベル賞の検索結果をもとに，手で時情報を抽出し作成した年表の一部分である．記述されていた時情報の位置に印が打たれ，その時に何らかの検索キーワードに関連する出来事が起きたことのみが分かる．Google Mapsの図2.3と比べると，年表表示の場

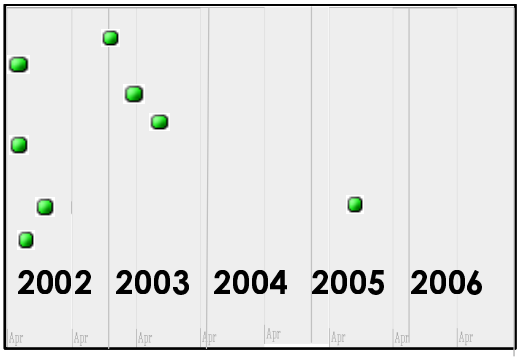

図 3.1: ノーベル賞の検索結果から抽出した5件のコンテンツを年表上に配置

合，印を付けただけでは有用性が低いことが分かる．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2.3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出力結果から分かる．一方，年表上の固有の位置に印が付いた場合，これはキーワードとその「時」を表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りそれ以上を年表から読み取ることはできない．これでは，筆者が目標とする，利用者にとって新たな発見や知識の創発を促す効果を期待することはできない．

これの解決策として，印に情報を補足することでその「時」の出現した理由を理解する助けとすることを考えた．利用しやすい情報として，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を表示する手法を実施した．これを図3.2に示す．出力に利用している検索結果コンテンツは，図3.1と同一である．(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ことが分かる．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，任意の「時」を補足する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．

そこで筆者は，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”ヘッドライン”を用いることを考えた．これを図3.3に示す．これも，出力に利用している検索結果コンテンツは図3.1と同一である．時情報の位置にその情報を添

図 3.2: 左:抽出した時情報に Title タグ情報を付与　右:URL 情報を付与

図 3.3: 時情報に時情報とキーワードの関係を示す用語を付与

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．

3.3 本章のまとめ

本章では，利用者に知的創発を促す年表俯瞰実現のために解決しなければならない課題を特定した．

一点目は「年表に出力する最適な時情報の抽出」である．不規則な形式で記述された時情報を含む検索結果コンテンツ群から，年表に利用可能な日付情報を抽出することの難しさと必要性を述べた．また利用者の混乱を避けるために，莫大な時情報の中から，利用者に発見をもたらす出来事を説明するもののみを抽出・配置することの必要性を述べた．

二点目は，キーワードとその時情報の関係を示す “ヘッドライン” の生成である．年表では印を打つだけでは有用性が低いことを示し,「時」とキーワードの関係を示すヘッ

ドラインが，年表の効果を最大化することを述べた.

第4章　提案手法

4.1　年表に配置する文の適正

検索結果コンテンツ群には，様々な時情報と，その「時」に関連する情報が記載されている．その中には，年表に表示すると利用者にとって有用な新たな発見を与える可能性が高いものと，そうでない (年表上に表示しても無意味な) ものが含まれている．

前者の例を 4.1に示す．これらは,「ノーベル賞」の検索キーワードを用いた．検索結

(検索結果コンテンツ C1 内の文 1)
下村脩さんは 1960 年にアメリカに渡り、発光生物が光を発するしくみを研究しました
(検索結果コンテンツ C2 内の文 1)
日本は 1949 年に初めて物理学賞を受賞して以来、物理学で 7 人の受賞者を輩出した
(検索結果コンテンツ C3 内の文)
ローレンス・ブラッグ 1915 年 25 歳で物理学賞受賞

図 4.1: 年表上に表示した時に有用性の高い文の例

果コンテンツ群から 1 つずつ時情報を含む文を抜き出した．例には，登場人物や関連する組織，場所，事柄などが含まれており,「時」と検索キーワードの関係となる出来事を十分に説明する情報が含まれている．これらの特徴的な要素を使うことで，検索結果コンテンツの情報編纂の有効性が見込めると考えた．

一方，年表表示に無意味な例を，図 4.2に示す．これらは，図 4.1と同様に「ノーベ

(コンテンツ C1 内の文 2)
ついに 2008 年の受賞者が発表された
(コンテンツ C2 内の文 2)
投稿日:2008/05/15 17:54:01
(コンテンツ C6 内の文)
2008/08/12 更新

図 4.2: 年表上に表示した時に有用性の低い文章の例

ル賞」での検索結果コンテンツ群に含まれる文である．一見して，年表に表示しても利用者の気付きや発見に繋がると判断しにくい．

そこで，筆者は年表に表示するコンテンツの情報を選択することにした．この時，情報中の固有名詞の存在に注目した．固有名詞は，名詞の一種であり，それ以外には存在しない特定の対象を表す名詞である．具体的には人名，地名，組織名などが挙げられる．図 4.1に照らすと，固有名詞として「下村脩」「アメリカ」「アトランタ」「マクドナルド」「シーシェパード」など，情報の核となる特徴的な単語として特定される．図 4.2を参照すると，それら固有名詞は特定されない．よって，年表に表示する情報は，**固有名詞を所有する文に限定する**．

4.2　時情報の抽出：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．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 つが欠落するときは，それぞれの状況に応じて細くルールを起動し，年，月，日からなる情報を生成する．具体的には以下の通りである．

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

これらのルールを用いて，検索結果コンテンツから年表に配置可能な時情報を抽出していく．

ルール 1 は，年月日の情報が全て含まれており，そのまま年表上に配置できる．ルール 2 は，年の情報を含み，年表上に配置する位置をある程度決定できる時情報を対象としたルールである．このルールでは，情報が不足している時情報を補完することで，年表表示に配置可能なものとして利用する．ルール 3 は，年の情報が不足している日付情報である．この形式の記述はニュースサイトに多い．このパタンの時情報は，年表配置に必要不可欠な年情報をサーバの発信情報から補完し，年表配置に利用する．

4.3　ヘッドラインの生成：時に関係する重要語の組

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．そこで、できるだけ正確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 (形態素解析と構文解析) 結果と係り受け情報を利用し，時情報を含んだ一文から時に関する重要語を数語抽出，列挙する手法を用いた．

4.3.1　重要語の抽出

時情報と検索キーワードの関係を表現する要素となる重要語には次の 5 種類ある．これらをどのように組み合わせるかは次節 4.3.2に示す．

固有名詞

固有名詞は，名詞の一種であり，それ以外には存在しない特定の対象を表す単語である．人名，地名，組織名などが挙げられる．これらは，情報共有に重要とされている 5W1H の Who(誰が)，Where(どこで) を表す単語であり，文意の特徴を掴むために重要な役割を果たす．固有名詞は，形態素解析を行う際に用意した辞書ファイルを参照することで，特定することができる．具体的な例を，図 4.3に示す．これは，時を含む一文から，人名と場所を抽出している例である．

括弧で特定されている文言

文章には,「」や『』、【】に囲まれた部分が存在することがある．これらは，一般的にタイトルや名称など特に注意を喚起したい語句を挿入する場合に用いられることが多い．従って，特徴的な語として活用できる．

具体的な例を，図 4.4に示す．これは,「」内の語を抽出する例である．

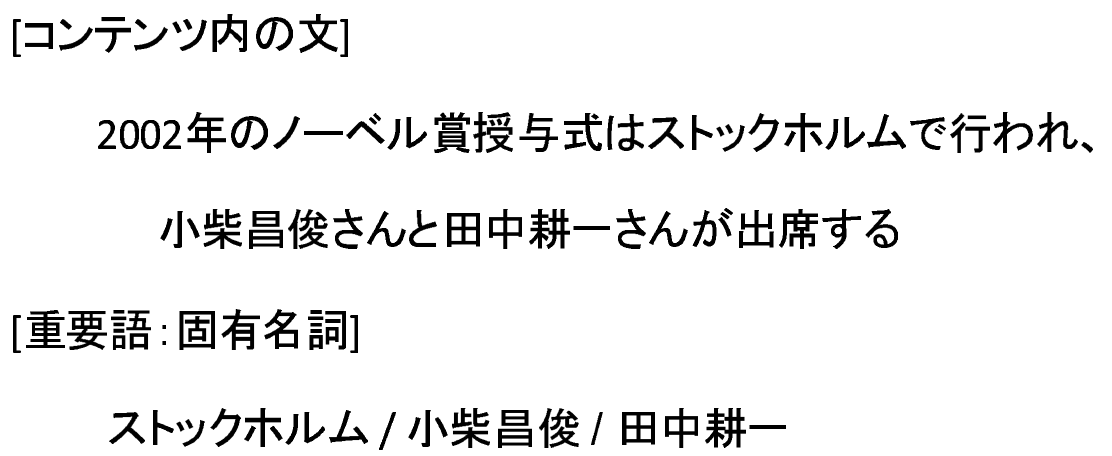

図 4.3: 固有名詞の抽出

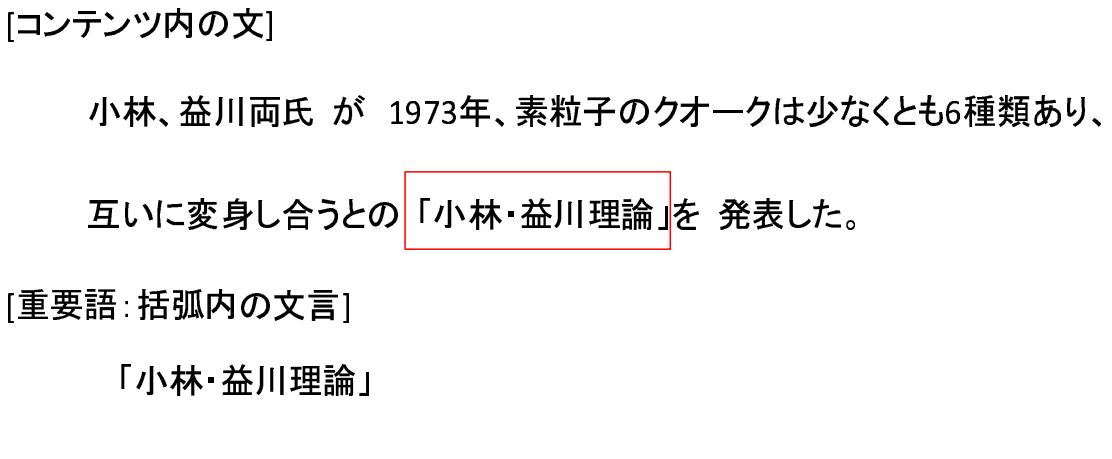

図 4.4: 括弧内の文言の抽出

程度副詞とその前後の名詞

副詞とは，用言を修飾し，動作・作用の状態や程度を詳しく表わす単語である．文中に副詞が登場する例を挙げると,「初めて」受賞した,「ことごとく」失敗した，など副詞は文を特徴付ける単語となっていることが多い．よって，副詞とその前後の単語は重要語とみなせる．副詞全てを対象とした場合，特徴的でないもの (「ゆっくり」「うっかり」など) を含んでしまう可能性があるため，質や状態などの程度を表わす程度副詞のみに限定する．具体的な例を図 4.5に示す．

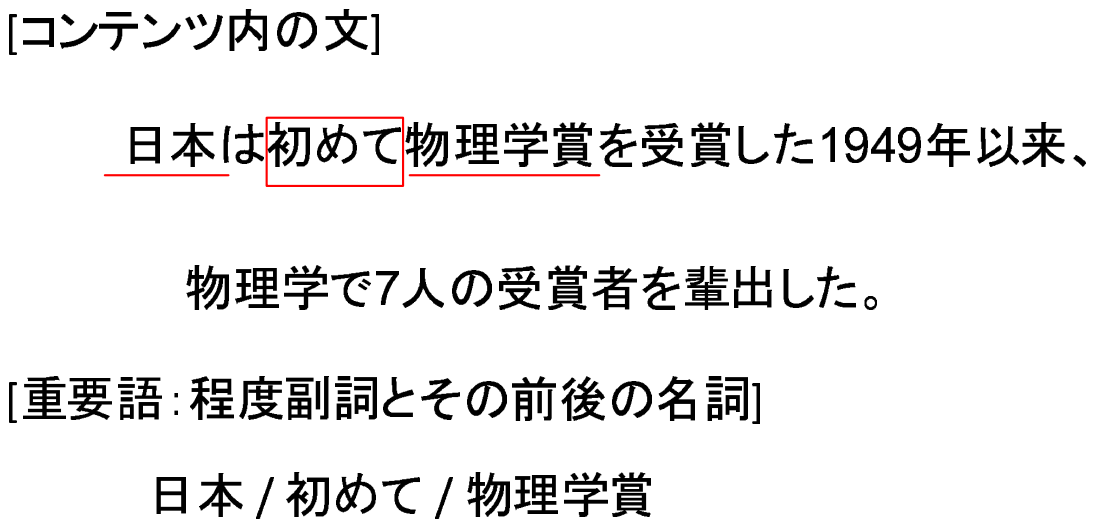

図 4.5: 程度副詞とその前後の名詞の抽出

「時」の係り先の用言と係り元の名詞

「時」の係り先の用言は，その時に発生したイベント (出来事) を直接表現し，他の用言から明確に区別することが可能である．さらにその用言の係り元名詞を用いると，より的確に事実を示せる．具体例を図 4.6に示す．図 4.6は，時の係り先の用言と，それに係る最初の名詞利用することで，その間を「時に関係する箇所」と特定している．さらにこれを応用して，その用言の係り元名詞を用いると，より的確に事実を示せる．これにより，時と関係する箇所の「何が」「どうした」の 2 つが明確になると考えた．具体例を図 4.6に示す．

「時」の前後の単語 (名詞)

時が用言以外の単語に係っている場合,「1900 年の東京」や「2008 年の北京オリンピック」などほとんどの場合隣接する単語を修飾している．これら時の隣接名詞も，時と関連が深いと考えた．

4.3.2　ヘッドライン生成ルールと適用順序

ヘッドライン生成アルゴリズムの基本的な考え方は，4.1を踏襲する．つまり，固有名詞の利用を原則とし，それ以外の重要語との組み合わせによってヘッドラインを生成する．

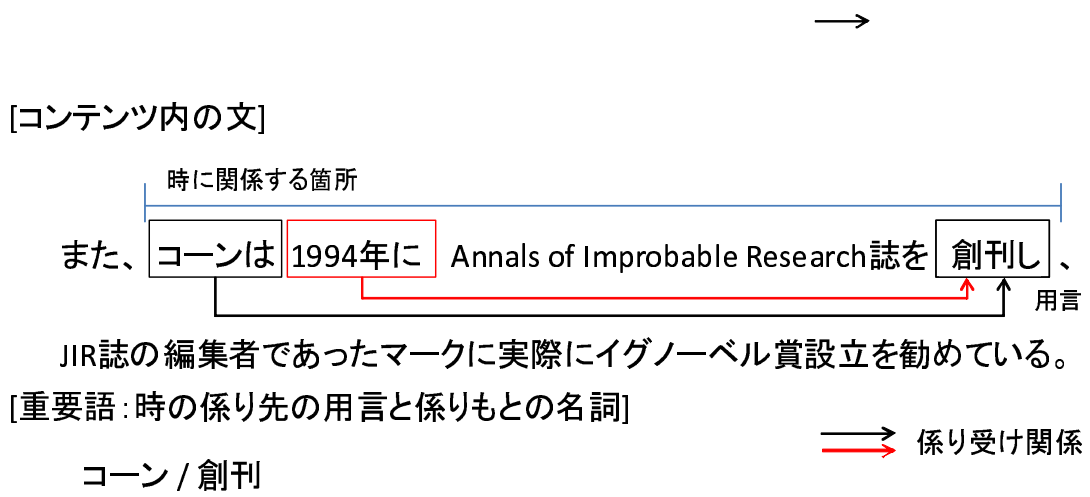

図 4.6: 時の係り先の用言とその関連語を抽出

1. ヘッドライン生成ルール 1
   4.2節で示した時表現を持つ文を形態素解析した結果
   括弧内の文言と括弧内から括弧外へ係る用言，用言の主語を要素に持つ時，この 3 つをスラッシュ (/) で結合する．

2. ヘッドライン生成ルール 2
   4.2節で示した時表現を持つ文を形態素解析した結果
   程度副詞と副詞の直前の名詞，副詞の直後の名詞を要素に持つ時，この 3 つをスラッシュ (/) で結合する．

3. ヘッドライン生成ルール 3
   4.2節で示した時表現を持つ文を形態素解析した結果
   「時」の係り先の用言，用言の主語，用言の前の名詞，さらにその直前の名詞を要素に持つ時，この 4 つをスラッシュ (/) で結合する．

4. ヘッドライン生成ルール 4
   4.2節で示した時表現を持つ文を形態素解析した結果
   「時」の直前の名詞,「時」の直後の名詞，さらに直後の名詞を要素に持つ時，この 3 つをスラッシュ (/) で結合する．

一方，検索結果コンテンツ内の「時」を含む文は，これらのルールの複数に適合し，その結果複数のヘッドラインが生成される．筆者は，年表上に配置したときの有用性を最大にするために，ルールの適用順を検討した．図 4.7に示す．

一般に括弧内の文言は 1 つの要素として扱われることが多く，分解しても有用性が上がることは少ない．そこでルール 1 を最優先に適用し，ヘッドラインが生成された文は他のルールを適用しない．次に，程度副詞が文中に存在するかどうかを判定する．程度副詞は，前後の名詞の特徴量を増加させる役割を持っている．そこで全ての名詞

の特徴量を同一に扱う他の 2 つに優先してこのルール 2 を適用する．次に，出現する要素の特徴量に (程度副詞による) 重み付けが無い場合，動詞 (用言) がより重要であるという先行研究 [14] を踏襲する．基本的にはルール 3 を適用する．さらにこの時，用言を修飾する可能性の高い，用言の前の名詞を特定することで用言の特徴を増加できると考え，それをルールに追加した．最後に残った文に対して，ルール 4 を適用する．

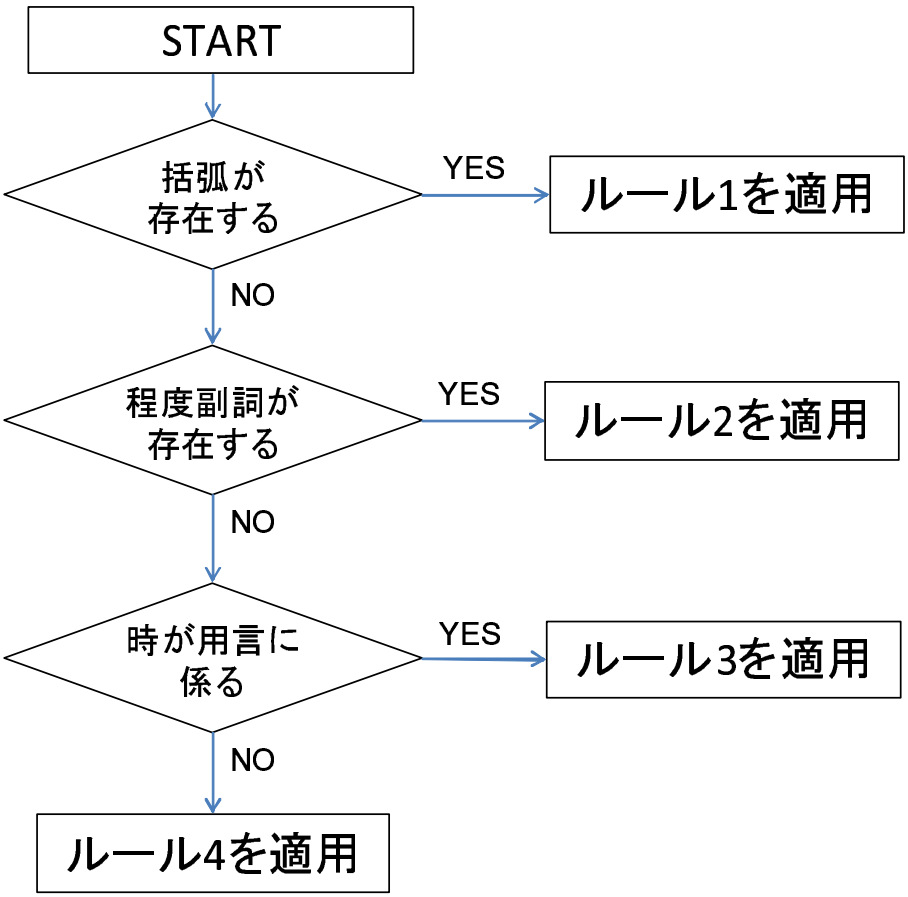

図 4.7: ヘッドライン生成プロセス

4.4 本章のまとめ

本章では，年表表示に必要な “時情報” と，その「時」と検索キーワードの関係を表現する “ヘッドライン” を検索結果コンテンツ群から抽出，生成する具体的な手法について述べた．

時情報抽出では，複数の時情報記述パターンを定義し，それにマッチした文字列を年表表示に利用する．

ヘッドライン生成では，できるだけ短い字数で構成させるという制限から，重要語を抽出し，組み合わせて表示するという手法を考案し，ルールを定義した．形態素解

析と構文解析の技術を用い，固有名詞抽出や係り受け関係を利用することで時に関係する重要語を抽出，年表上への配置を実現する．

第5章　検索結果年表表示システム概要

5.1　システム構成: 汎用検索エンジンの拡張

本サービスは，検索エンジンとは別個ではなく，汎用検索エンジンを拡張する形で実現する．

標準的な検索エンジンは [1] の示す通りであり，Google[2]，Yahoo![3] 等も同一の構成をとる．各機能について説明する．Gathering は，Web を巡回しコンテンツを収集する機能である．Extracting は，収集したコンテンツを保存する機能である．Indexing は，保存されたコンテンツを高速に検索するためにコンテンツ中の単語リストとコンテンツ ID を生成し，その二つを紐付け保存する機能である．Querying は，ユーザからの検索リクエストを受け取る機能である．検索キーワードによる検索の他には，ファイルタイプやドメインの指定などが存在する．Ranking は，検索結果に出力されるコンテンツリストにおけるコンテンツの優先順序を決定する機能である．Presenting は，検索結果を利用者に提供する機能である．

図 5.1は，この標準アーキテクチャに、筆者が提案する検索結果コンテンツの年表表示を付加する場合の新構成である．前述のモジュールにさらに 3 つのモジュール (「編纂用の「時」 Extracting」,「編纂用 Headline Generating」」,「Tagging」) を追加する．「編纂用の「時」 Extracting」とは，Indexing されたコンテンツにから，年表表示可能な時情報を抽出する機能である.「編纂用 Headline Generating」とは，抽出された時情報を含む文からヘッドラインを生成する機能である.「Tagging」とは，上記の 2 つの機能で作り出された情報を，indexes 中の各コンテンツに対し拡張タグデータとして記録する機能である．

5.2　処理フロー概要: 検索エンジン・indexes を利用

前節で示したシステム構成上で，検索結果の年表表示実現のための情報処理は次のように実施する．

定期的に，検索用の indexes からドキュメントを収集し，それに対して時情報抽出ルールとヘッドライン生成ルールを適用する．抽出・生成されたデータは，indexes に存在するドキュメントの拡張タグ部分に追加され，更新される．そして利用者から年表表示の呼び出しがあった時，表示機能は indexes から検索結果を取得し，その拡張タ

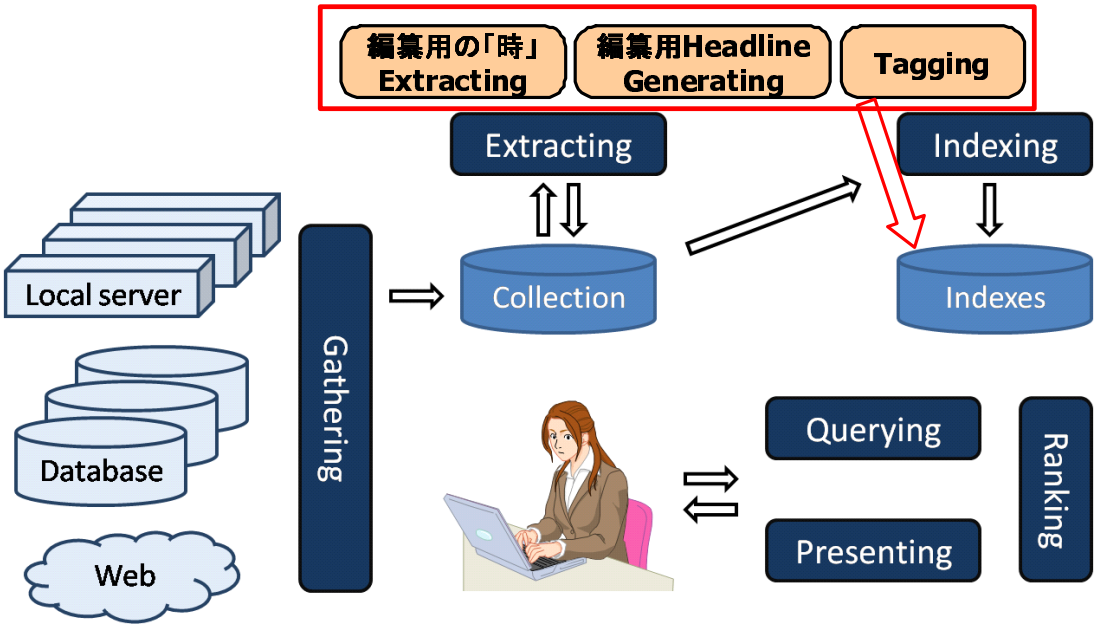

図 5.1: 既存検索エンジンの拡張 - システム構成図 [1]

グ部分にある時情報とヘッドラインを用いて年表を生成する．こうすることで，システムは通常の検索結果出力と同じように年表用のデータを提供することができる．

本システムの処理フローについて図 5.2を用いて説明する．

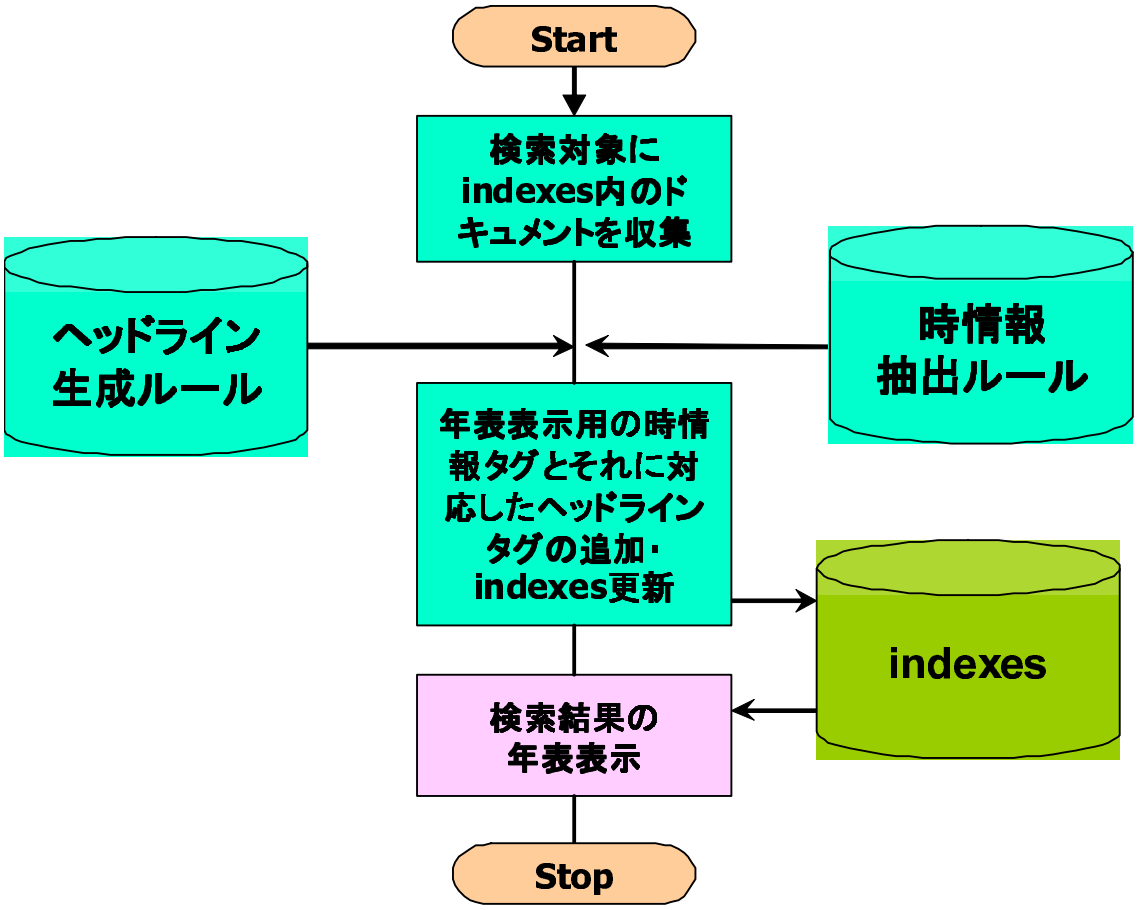

図 5.2: 年表表示機能の情報処理フロー

第6章　有効性の効果検証

6.1　抽出した時情報の有用性

第 4.2節で示したパターンマッチを用いてコンテンツから時情報を抽出する手法の，精度 (Precision)・再現率 (Recall) を検証する．精度は，抽出した情報の中に，年表に利用可能な時情報 (年，月，日もしくは年及び年，月で構成されているもの) がどれだけ含まれているかの割合とし，再現率は，コンテンツ群のすべての時情報から，漏れなくそれらを抽出する割合とした．

実験は,「ノーベル賞」「iPS 細胞」「クラゲ」「ロス疑惑」の 4 つのキーワードで検索し，それぞれ上位 100 件のコンテンツ群 (Data1-Data4 とする) に対して行った．それぞれのコンテンツ群を手動で確認して抽出した正解と，パタンによって抽出した時情報を比較する．

結果を表 6.2に示す．精度を見ると，平均 96.7%の時情報が抽出できており，これは，

	Data1 ノーベル賞	Data2 iPS 細胞	Data3 クラゲ	Data4 ロス疑惑	平均
正解「時」情報件数	1884	890	563	850	
精度	98.9%	98.9%	93.9%	95.2%	96.7%
再現率	75.7%	93.4%	82.2%	88.2%	84.9%

表 6.1: 時情報のパタン抽出:精度再現率

抽出した情報はいずれのテストケースでも 90%以上が年表表示に利用可能な時情報であったことが分かる．誤って抽出してしまった代表的な例として，分数表現 (「1/3」三分の一) や，電話番号 (「02xx-45-89xx」) が挙げられる．再現率を見ると，キーワードによって若干ばらつきがあるが，平均すると 84.9%であった．これは，本手法がコンテンツ群に含まれる時情報の 8 割以上を網羅的に抽出できていることを示す．

6.2　生成したヘッドラインの有用性

生成したヘッドラインが，時情報とキーワードの関係を正確に示せているか検証する．検証は，検索結果コンテンツ群から抽出した時情報を含んだ文を対象に，手動で

生成した正解例と，本手法を用いて生成したヘッドラインを比較することで行う．有効性検証のテストは以下の手順で行った．

1. 任意のキーワードを元に検索結果コンテンツ群を収集
2. コンテンツ群から，時を含む文を抽出
3. 抽出文から，手動で重要語の組によるヘッドラインを生成
   (この正解例データセットを A とする)
4. 抽出文から，本手法を用いてヘッドラインを生成 (このデータセットを B とする)
5. B の，A に対する網羅率を計測

ここで網羅率は，正解であるヘッドラインおよびヘッドラインと他の語を出力した場合を網羅したと判定した．例えば、正解であるヘッドラインが語 (a,b,c) から成る時，提案手法で語 (a,b,c,x) が出力された時、これも網羅できた事象として判定した．一方，語 (a,c) が出力された場合，網羅されていない事象として判定した．これらは利用者が年表上に出力された結果を見て，より意味ある理解を行えると注目したことによる．

実験は「ノーベル賞」「オリンピック」「早稲田大学」「クラゲ 発見」「ボランティア」「鯨」の 6 つのキーワードで検索し，それぞれ上位 200 件のコンテンツ群を Data5-Data10 とした．ヘッドラインの正解例と本提案生成手法によるヘッドラインが同一かどうかを比較した．

	Data5 ノーベル賞	Data6 オリンピック	Data7 早稲田大学	Data8 クラゲ 発見	Data9 ボランティア	Data10 鯨	平均
正解事象数	43	45	28	22	34	40	
網羅率	87.8%	85.6%	80.3%	91.7%	89.0%	92.1%	87.6%

表 6.2: ヘッドライン生成:正解における網羅率

次に，全体の約 88%を網羅する適用ルールの内訳をグラフ 6.1に示す．適用されたルールは「時」の係り先や前後の単語を活用するルール 3 と 4 が，それぞれ全体の 1/3 ずつを占める結果となった．

6.2.1 先行手法との比較

形態素解析技術による時情報抽出と本パタンマッチの比較

比較対象として，形態素解析技術による時情報抽出を用いる．形態素解析のソフトウェアには，Chasen[15] を利用した．

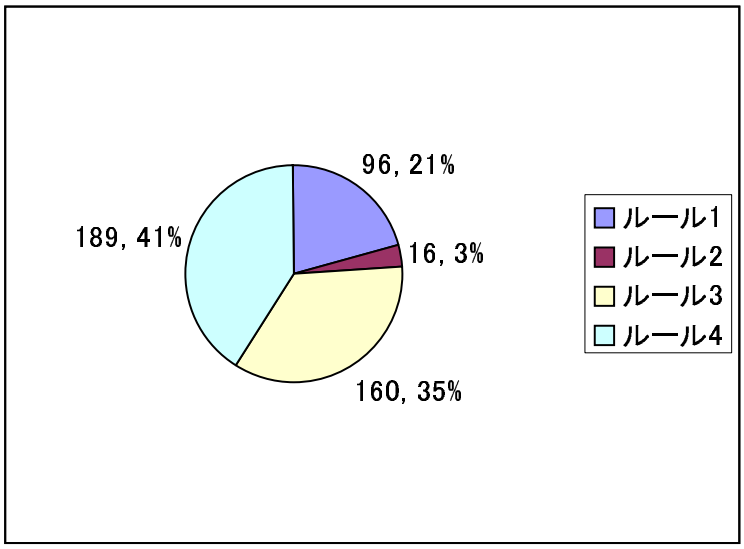

図 6.1: 適用ルールの内訳

6.1節と同様に Chasen の時情報抽出の精度・再現率を測定し，比較する．6.1節と同じコンテンツ群に対し，Chasen が月日を表す情報 (DATE) と解析した文字列が，年 (可能であれば月日も) を含む情報か判定し，数を測定した．

まず，Chasen の時情報抽出結果の数との比較を行う．実験結果を 6.2に示す．これは，

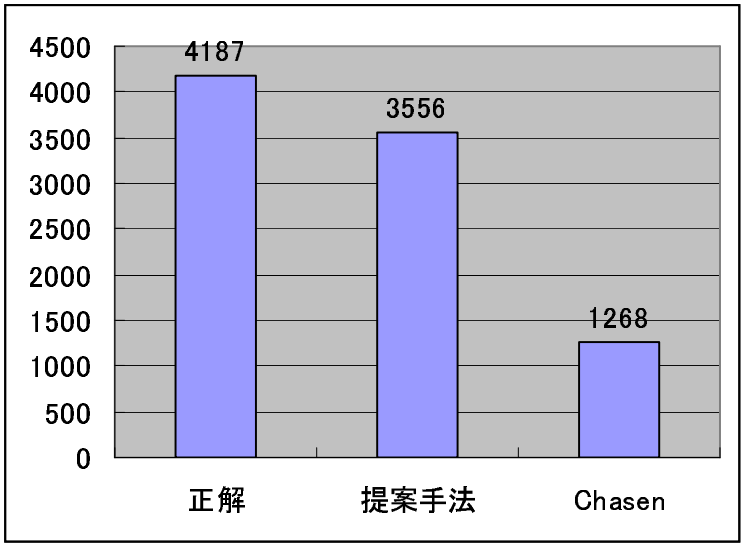

図 6.2: 提案手法と Chasen の抽出数比較

コンテンツ群中の年表表示できる時情報 (これを正解とする) と，本手法で抽出した時情報，Chasen で抽出した時情報の 3 つの数を表したものである．正解例が 4187 個存在

する中で，本手法は 3556 個，Chasen は 1268 個を抽出できており,，本手法は Chasen と比べ約 2.8 倍の時情報を正確に抽出できたと言える．

次に，抽出した時情報の精度を測定した．実験結果を図 6.3に示す．これは，本手法

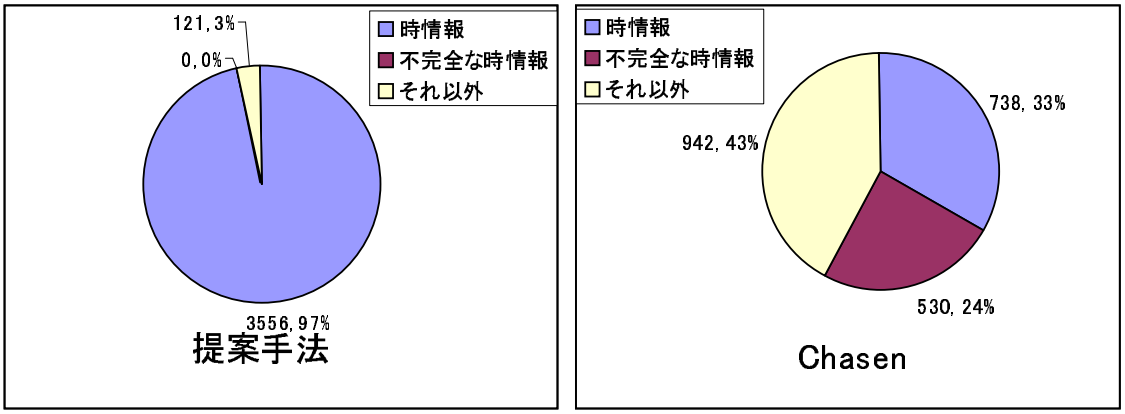

図 6.3: 提案手法と Chasen の精度比較

と Cabocha を用いてそれぞれ抽出した情報に，年 (可能であれば月日も) が含まれているか判定し，数を測定したものである．不完全な時情報とは，時情報の一部分のみが抽出できたものである．例えば，文の構成によっては,「2008 年 10 月」のうち「2008 年」のみを時情報としてしまうことが確認された．グラフより，本手法は Chasen に比べ約 1.68 倍の精度で年表表示に利用可能な時情報を抽出できると言える．

最後に，これらの結果から，本手法と Cabocha の精度再現率を 6.3にまとめる．

	提案手法	Cabocha
精度	96.7%	57.4%
再現率	84.9%	30.7%

表 6.3: 提案手法と Cabocha の精度再現率

第7章　考察

7.1　時情報抽出について

筆者は，年表表示に必要不可欠な時情報をコンテンツ群から抽出する手法として年(可能であれば月日も) を特定できるパターンを定義し，それを用いて抽出をおこなった [4.2].

第 6章の効果検証を参照すると，本提案パタンでの抽出精度は 96.7%であった．これは，コンテンツ群に存在する，年表表示可能な「時」をほとんど抽出できたと言える．年表上に表示できる時情報は，時の一点を指す絶対的なものでなければならないが，当初の予想では,「5 年前」や「20 年ぶり」といった相対的な年表記が多く含まれてしまうのではないかと懸念していた．実際に検証実験の結果を参照すると，それらは予想より大幅に少なかった．その代わりに，精度を下げる原因として新たに発見できたのは，分数表現と電話番号表記であった．分数表現は,「mm/dd」(月/日) の形式に対応するパターンが当てはまってしまうために，時情報として抽出されてしまう．このパターンは様々なコンテンツで頻出するために，ルールから除外してしまうと再現率が大幅に低下する恐れがあるため，除外できない．そのため，別の解決方法を模索していかなければならない．電話番号表記は,「yyyy-mm-dd」(年-月-日) の形式に対応するパターンが当てはまってしまう．この問題に関しては，yyyy の具体的な数値を厳密に設定することで対処できると考えている．年，月，日と 3 つの要素を揃えるこのパタンでは，年の情報はここ数百年のものであることが多い．コンテンツ群を確認しても西暦 1000 年以前の時情報の中に，このパタン形式で月日まで書き込まれているものは存在しなかった.(例えば 645-6-12 など) さらに，何千年も未来の西暦が登場することも確認できなかったため，yyyy の数値を 1000-3000 の間に設定することで，電話番号の抽出は防げるのではないかと考えている．電話番号の市外局番リストを参照すると，4 桁の 1 もしくは 2 で始まるものは北海道の一部地域のみであり，この対処は効果的に作用すると考えられる．

形態素解析技術 (Chasen) との数の比較では，正解の 4187 に対し，本提案手法では 3556，Chasen では 1286 という結果であった．Chasen の出力例を見てみると，時情報と特定しているものは「今日」や「明日」などの時を表す単語と,「年」「月」「日」という表記を含むものが大半であった．web 上には，本提案手法で示した「yyyy/mm/dd」や「yyy-mm-dd」などのパターンが時情報として頻出するため，これらを取りこぼしてしまった結果，抽出数が少なくなってしまったと考えられる．

7.2 ヘッドライン生成について

筆者は，時情報を含む文中の「時」とキーワードの関係を年表上に示すために，重要語の組によるヘッドライン生成手法を提案した.

ヘッドライン生成には当初，要約技術 (厳密には「一文要約」と「文圧縮」) の応用を検討していた. しかし，実際に web コンテンツに対して要約技術 [16][17][14] を適用してみると，それは適切でないことが分かった. web コンテンツ中には，様々な利用者によって自由に文章を書き込まれるものが多くある. それらのコンテンツを対象にした場合，日本語文法に照らすと不正確な文章が多く含まれてしまう. 要約技術では，形態素解析や構文解析が高度に駆使され要約文の生成がおこなわれるが，ニュース記事や論文のような「正しい日本語」で表記されることが保証されていない web コンテンツを対象にした時，期待通りの要約は実現できなかった. また，致命的であったのが，検索結果コンテンツは動的に変化するものであるため，特徴的・重要とみなされる単語を予め特定する tf-idf[18] を利用した手法 [14] が適用できないことであった. tf-idf とは，静的な大量のデータを対象に，単語を多くのドキュメントに出現する語（一般的な語) と特定のドキュメントにしか出現しないものに選り分け，特徴の強さをスコアリングする技術である. さらに，要約「文」を生成する手法では，どうしてもヘッドラインが長くなってしまい，直感的に理解しづらい，年表上に表示できない文字が多くなってしまう，という問題があった. これらの理由により，既存の要約技術はヘッドライン生成に適用できなかった.

本提案手法は，時に関係した重要語を抽出し結合するものである. 第 6章の効果検証を参照すると，平均 87.6%の文で正解の単語を網羅できていた. 提案手法の利点は，特徴的な単語や，時に直接関係する単語を抽出できる点である. こうした単語を列挙することにより，利用者は「時」とキーワードの関係示す出来事を端的に年表上で確認できる. また，重要な単語のみを結合させることで，助詞や接続詞など必要でない単語が削除され，年表の空間を有効に使うことができる.

第8章　結論

本研究は，検索結果の新表示方法の提案として「年表」 に注目し，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．そして検索結果コンテンツ群を年表上に編纂することで利用者に知的創発を促すことを目指した．

検索結果コンテンツの俯瞰に年表を利用することで，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．

本研究では，実用性のある年表を生成するために (1) 年表に出力する最適な時情報の抽出と (2) 効果的な「ヘッドライン」 生成の，2 つの手法を開発した．(1) を解決するための提案手法は，パターンマッチを用いた時情報の抽出であった．様々な形式で記述される時情報を正確に網羅するために，10 の抽出パターンを定義し，年表上に利用できる，年 (可能であれば更に月日) を持つ時情報を抽出した．次に，(2) を解決するために，形態素解析・構文解析の結果から「時」に関係した重要語を抽出・結合しヘッドラインを生成する手法を提案した．

手法の有用性を検証する実験を行った．提案した時情報抽出手法は，正解のデータに対し 96.7%の精度と 84.9%の再現率を達成できた．また，提案したヘッドライン生成手法は，正解として用意したヘッドラインを構成する単語を，87.6%の確率で網羅できた．これらの実験により，本手法による時情報抽出とヘッドライン生成手法の組み合わせは，検索結果コンテンツの年表を使った編纂に有効であると分かった．

謝辞

本論文の執筆にあたり，本当に多くの方々にお世話になりました．この場をお借りして，お礼を述べさせていただきたいと思います．

まず、本研究を進めるにあたり，ご指導をいただきました慶應義塾大学 環境情報学部教授 村井純博士に感謝します．修論や今後の研究の進め方について貴重な意見を頂きました．ありがとうございました．

同学部教授 中村修博士に感謝します．研究室の授業である研究発表の場で，研究に対する姿勢，進め方をご指導頂きました。

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准教授 嶋津恵子博士に感謝いたします．休日を問わず，親身に細かく指導して頂き，本当にありがとうございました．

慶應義塾大学村井研究室の同期であり、共に執筆に取り組んだ空閑洋平 (sora) 氏，金井瑛 (kanai) 氏に感謝致します．お二人がいなければ，僕の研究室生活はこんな楽しいものにはなりませんでした．一つ上の先輩である，水谷正慶 (mizutani) 氏に感謝致します．お兄さん的存在として，気軽に親身に相談に乗って頂きました．

また SING/IA*グループ時代に研究の基礎をご指導いただいた南政樹氏，小原泰弘氏に感謝致します．シスコシステムズ 小椋康平 (koh39) 氏から伝えていただいたゲームの文化は，最高の時間をもたらしてくれました．ありがとうございました．後輩である佐藤龍 (tatsurou) 氏，波多野敏明 (haeena) 氏には，研究室や授業など，あらゆる面でご協力いただきました．ありがとうございました．いつも遊び相手になってくれる，NTT コミュニケーションズ 尾崎隆亮 (tacos) 氏，ありがとうございます．

研究をご指導くださった研究室の先生方，斉藤 賢爾氏，武田 圭史氏，楠本博之氏，バンミーター・ロドニー氏，重近 範行氏，吉藤英明氏，朝枝 仁氏，植原啓介氏，三次仁博士，鈴木茂哉氏，稲葉達也氏，中根雅文氏，土本康生氏，Achmad Basuki 氏，に感謝致します．

最後に私の大学生活を全力で支えて下さった，父 睦夫，母 和子，兄 保之，和明に最大限の感謝の意を述べたいと思います．

皆様，本当にありがとうございました．

2009 年 1 月
奥村 祐介

参考文献

[1] David Hawking. Web search engines: Part 1. *IEEE Computer*, 39(6):86–88, 2006.

[2] Google Inc. Google search. `http://www.google.com`, 2008.

[3] Yahoo Inc. Yahoo! japan. `http://www.yahoo.co.jp`, 2008.

[4] iProspect. iprospect search engine user behavior stury. `http://www.iprospect.com/premiumPDFs/WhitePaper_2006_SearchEngineUserBehavior.pdf`, 2006.

[5] M. T. Hansen, N. Nohria, and T. Tierney. What's your strategy for managing knowledge? *Harvard Business Review*, 77(2), 1999.

[6] K. Shimazu, T. Arisawa, and I. Saito. Interdisciplinary contents management using 5w1h interface for metadata. *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*, pages 909–912, Dec. 2006.

[7] MIT Simile Project. Simile timeline. `http://code.google.com/p/simile-widgets`, 2008. http://code.google.com/p/simile-widgets, retrieved 2008-12-11.

[8] today's timetable. `http://www.cod3.net/timetable/`, 2008.

[9] Wikipedia, the free encyclopedia. `http://www.wikipedia.org`.

[10] Google Inc. Google maps. `http://maps.google.co.jp/`, 2008.

[11] Google Inc. Google alternate views for search result: Timeliveview. `http://google.com/views?q=%3Atimeline`, 2008.

[12] mediajam. ニュース年表. `http://mediajam.info/timeline/`, 2008.

[13] Allofme - beta. `http://www.allofme.com/`, 2008.

[14] 満美子 岡, 剛弘 小山, and 良寛 上田. 句表現要約の句合成手法. **情報処理学会研究報告**. **自然言語処理研究会報告**, 99(2):101–108, 19990120.

[15] 裕治 松本, 一馬 高岡, 正幸 浅原, and 拓 工藤. 茶筌と南瓜による日本語解析 : 構文情報を用いた文の役割分類 (¡シリーズ特集¿研究のツールボックス (1)). **人工知能学会誌**, 19(3):334–339, 20040501.

[16] Kevin Knight and Daniel Marcu. Statistics-based summarization - step one: Sentence compression. pages 703–710, 2000.

[17] 嶋津恵子, 山根洋平, 門馬敦仁, 桜井哲志, and 古川康一. 相関ルールの活用によるコールセンター業務への貢献. **第** *65* **回情報処理学会 全国大会**, 200303.

[18] Gerard Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.