第1章　序論

概要

本章では，本研究の背景・目的および論文の構成について示す.

1.1. 研究の背景・動機

情報やデジタルコンテンツの短命化が進んできている．その理由として，アクセス可能な情報量の急激な増大と情報伝播速度の速いサービスの一般化があげられる．特に Web 上の情報は，急速に増加しているとともに，玉石混交化が進んできていると言われている [?]．また，例えば Twitter[?] などのマイクロブログをはじめとした情報伝播速度が速いサービスも一般的に利用されるようになってきた [?]．このようなサービスの一般化やネットワーク速度の向上などによって，情報が流行する速度，さらには情報が廃れる速度も速くなってきた．例えば，Twitter 情報の中には，急速に流行して短い期間に多くの人から見られるが，急速に廃れて 1〜2 週間経過するとほとんど見られなくなるような場合も多い．さらに，このような特性のある Twitter やさまざまなソーシャルメディアによって急速に広まる Web ページも増加してきているが，急速に広まるということは急速に廃れる可能性が高い．こういったことから，利用される期間の短い情報やデジタルコンテンツがだんだん増加してきているということが言える．さらに、Web 以外に関しても、製品などのライフサイクルの短命化 [?] が起こっている。

このように短期的にしか利用しないような情報が増えてきたため，これまでの情報検索システムなどでは，長期的に利用可能な情報を見つけることが難しくなってきた．これまでのシステムでは，特に長期的な情報に焦点をあてていないため，短期的な情報の割合が増加した場合，短期的な情報が提示される確率があがる．さらに、短期的な情報でも一時的に人気や流行となる場合があり、その場合多くのシステムやサービスによって提示される。こういった流行や人気の情報は、多くのメディアやソーシャルメディアで取り上げられることが多いため、人々の目に触れる機会は多い。このような背景に対して，長期的に利用する情報に特化して取得するようなシステムはほとんど存在しない．そこで本研究では，長期的に利用できる情報へのアクセス方法に課題があるのではないかと考えた．

ところで，本や商品に関しては，ベストセラーとロングセラー [?] という指標がある．ロングセラーと長期的に利用されるということは,「売れる」と「利用される」の違いはあるものの，類似した指標であると考えられる．そこでロングセラーの有用性を例として，長期的に利用される情報の有用性について考えてみる．本や商品に関して言えば，ベストセラーを好む人もいれば， ロングセラーを好む人もいる．以上のような特徴がある一方，Web 上の情報はベストセラーといったような観点か，他人からの推薦といったような観点からの評価が主で，ロングセラーという観点での評価はほとんどされていない．Web の歴史はまだまだ浅く，商品の売れ方と Web では異なるところも多いが，Web 上の情報についてもベストセラーのような人気の高い情報を好む人もいれば，ロングセラーのように長期的に利用される情報を好む人もいる．筆者自身は，長期的に利用される情報に関心があるが，どちらの情報が欲しいかはユーザの好みだけでなく状況によっても変わってくる．そのため本研究では，商品だけでなく Web 上の情報に関しても，長期的に利用されているかどうかという指標も有用であると考えた．

そこで，本研究では，長期的に利用される情報という視点に着目し，こういった情報を見つけるための手法「長期的検索」を提唱する．このような情報を取得するためには，今までよく利用されていたリンク数，アクセス数，評価数，関連度などの指標だけではなく，時間情報を利用し，情報がどれだけの期間利用されてきたかという新しい評価指標が必要である．本研究では，この指標のことを「長期度」と呼ぶ．

1.2. 研究の目的

本研究では，主に以下の 3 点を目的とする．

- 検索者が長期的に利用できる情報を手軽に見つけられるようにすること
- 長期的に利用されている情報の有用性が高いことを示すこと
- 本研究で提案する長期的な情報を検索するシステムの有効性を示すこと

本研究の一番の目的は，人々が長期的に利用できる情報を手軽に見つけられるようにすることである．その理由は，長期的に利用できる情報は有用である可能性が高いにもかかわらず，見つけるための手法が確立されておらず，発見することが容易ではないからである．

次に，本研究では，長期的に利用されている情報が有用である可能性が高いことを示す必要がある．そのために，これまで長期的に利用されてきた情報は，今後も長期的に利用できるかどうかを調査する必要がある．今後も長期的に利用できるということは，有用性が高いということが言える．

さらに，実際に提案するシステムが有効であるかどうかを示す必要もある．このために，他の検索システムなどと比較して，提案システムの優れている点を明確にする必要がある．

1.3. 用語定義

本論文で使用するいくつかの用語に対して，用語の意味を説明する．以下に示すのは，本論文で利用する際の用語の意味である．

***1** マイクロブログ
主に 140 文字程度の短い文章を書いて Web に公開する短いブログ．代表的なマイクロブログサービスとしては，Twitter があげられる．ブログとの違いは，手軽である点とユーザ同士のコミュニケーションを支援するソーシャルネットワーク的な側面が強いところである．

***2** 長期度
その情報がどれだけ長期的に利用されているかを表す指標．

***3** 情報フィルタリング [?]
大量の情報の中から，ユーザにとって必要な情報を取り出し，不要な情報を除外する処理を自動的に行う技術のこと．本論文では，広義の意味として情報フィルタリングという用語を利用しており，情報の収集と排除双方の意味を含む．

***4** 情報レコメンデーション [?]
ユーザの興味や嗜好に応じて，お勧めの情報を提供すること．

***5** パーソナライズド検索
パーソナライズド検索とは，検索ユーザの過去の検索履歴などから，ユーザの興味や関心を推定し，それぞれのユーザに合わせた検索結果を提示する検索手法のことである．

***6** folksonomy
Web 上のデータにおいて，ユーザ自らが情報の分類・収集を行うこと．「人々」（folks）と「分類」（taxonomy）とを掛け合わせた造語である．例えば，ソーシャルブックマークなどでは，ユーザがブックマークしたページに自由にタグを付与できるが，このユーザ自身が自由にタグをつける行為も folksonomy の一つである．タグをつけることにより，タグを付けた情報にアクセスするための検索や分類に役立てることを目的としている．

***7** 体系だった知識を得られる Web ページ
ある分野に対して，基礎的な内容から応用的な内容までの幅広い知識を得られる Web ページ．また，ある分野に関する重要な知識がまとめて記述されているような Web ページ．

***8** RSS リーダー
指定した Web サイトの更新情報を一定時間ごとに自動的にダウンロードし，更新があると記事へのリンクを表示してユーザに知らせるツールのこと．

***9** ソーシャルブックマーク
インターネット上で自分のブックマークを不特定多数のユーザに公開し，有益な Web ページを共有する Web サービスのこと．

***10** スピアマンの順位相関係数
順位データから求められる相関の指標である．ノンパラメトリックな指標であり，2 つの変数の分布について何も仮定せずに，変数の間の関係を評価するものである．

***11** 適合率
検索結果として得られた文書中にどれだけ検索に適合した文書（適合文書）を含んでいるかという正確性の指標である．情報検索システムの評価を行う際に一般的に利用される指標であり，これとともに再現率が利用される場合が多い．適合率は以下の数式で表される．

$$\text{適合率} = \frac{R}{N} \tag{1.1}$$

R:検索された適合文書の数
N:検索結果の文書の数

***12** 再現率
検索対象としている文書の中で検索結果として適合している文書（適合文書）のうちで，どれだけの文書を検索できているかという網羅性の指標である．情報検索システムの評価を行う際に一般的に利用される指標であり，これとともに適合率が利用される場合が多い．再現率は以下の数式で表される．

$$\text{再現率} = \frac{R}{C} \tag{1.2}$$

R:検索された適合文書の数
C:全対象文書中の適合文書の数

***13** バースト
爆発する，破裂する，急に起きる，勢い良くでるなどの意味を持つ．

***14** Google サジェスト
Google 検索の補助機能で，ユーザが検索キーワードを入力しているときに，よく検索されるキーワードの候補を提示する機能である．

***15** Google Insights for Search
Google が提供している検索キーワードの検索ボリュームを分析することができるサービス

である．指定した検索キーワードに対して，Google での 1 週間毎の検索ボリュームを取得することができる．2004 年以降の検索ボリュームの時系列データを取得でき，また地域ごとのデータなどより詳細な分析も可能である．2013 年 1 月 10 日現在，このサービスは Google トレンドと合併し，Google トレンドにて同様の機能を利用することができる．

***16** Google トレンド
Google が提供している検索キーワードの検索ボリュームを分析することができるサービスである．Google Insights for Search とほぼ同等のサービスを提供している．複数のキーワードについて，検索ボリュームの比較をすることも可能である．

***17** スニペット
スニペットとは，一般的には「切れ端」「断片」といったような意味の英語であるが，IT 用語としては，検索エンジンの検索結果の一部として表示される Web ページの要約文のことを表す．

***18** カイ二乗検定
カイ二乗検定とは，帰無仮説が正しければ検定統計量がカイ二乗分布に従うような統計学的検定法の総称である．

1.4. 論文の構成

本論文の構成は，以下の通りである（図**??**）．

第 2 章　背景と関連領域

第 2 章では，本研究の背景と関連領域について述べる．

第 3 章　長期的情報検索の提案

第 3 章では，本研究で述べている「長期的な情報」の定義と「長期度」の定義を行う．また，長期的な情報を検索するための手法を提案する．

第 4 章　長期的な Web ページの検索

第 4 章では，長期的に利用されている Web ページを検索するシステム「セレクトブクマ」について述べる．

第 5 章　長期的な検索キーワードの提示

第 5 章では，長期的な検索キーワードを提示するシステム「MasteryEye」について述べる．

第 6 章　関連研究

第 6 章では，本研究に関連する研究領域について整理し，本研究の特徴や位置づけについて述べる．

第 7 章　考察と展望

第 7 章では，本研究の考察と展望について述べる．

第 8 章　結論

第 8 章では，本研究の成果をまとめ，本論文を総括する．

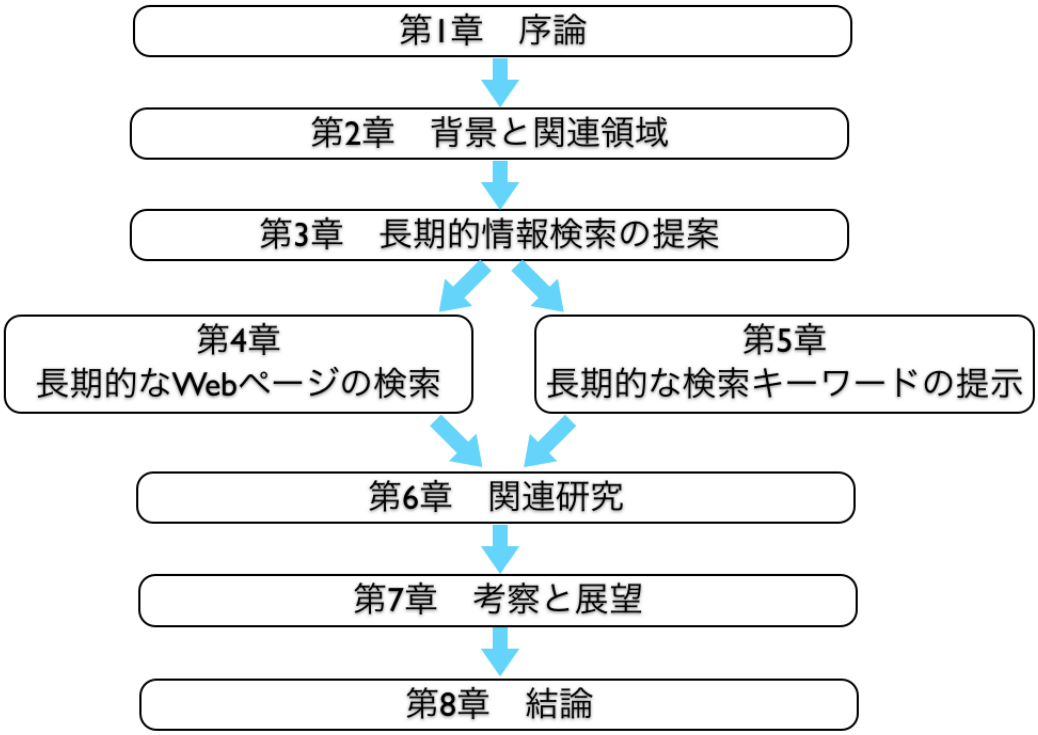

図 1.1: 本論文の構成