第１章　長期的情報検索の提案

概要

本章では、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、長期的検索を提案する。

1.1. 長期度

本研究では、長期的な情報を取得するために、「長期度」という指標を提案する。ここではその長期度について述べる。

1.1.1. 長期度とは

本研究では、その情報がどれだけ長期的に利用されてきたかを表す指標として、「長期度」という指標を定義する。長期度が高い程、長期的に利用されている情報、長期度が低い程、短期的にしか利用されていない情報とする。本研究では、長期的に利用されている情報を取得するシステムを実装するため、この長期度が情報を評価するための主な軸となる。現状では、長期度の計算手法は、情報の種類によって異なる手法を用いている。その理由として、情報の種類によって、情報の利用のされ方や一般的な利用数や利用される期間などが大きく異なることがあげられる。長期度の具体的な計算手法については、第 4 章〜第 6 章で述べる。

1.2. 長期的な情報

ここでは、本研究における長期的な情報の定義を示す。

1.2.1. 長期的な情報の定義

本研究で述べている長期的な情報とは、以下の 2 種類の意味を含む。

(1) これまでに長期的に利用されてきた情報

(2) 情報を取得した人が、今後長期的に利用できる情報

本研究では、(1) の情報は、(2) である可能性が高いという仮説を立てる。この仮説の元、(1) の情報を取得するシステムを開発し、システムの有効性と仮説の検証を行う。

1.2.2. 長期的な情報と短期的な情報の特徴

表 1.1 に、長期的な情報と短期的な情報の特徴をあげる。それぞれの情報は、以下のような傾向が強い。

表 1.1: 長期的な情報と短期的な情報の特徴

長期的な情報の特徴	短期的な情報の特徴
定番情報	流行情報
体系だった情報	断片的な情報
安定している	不安定である
じわじわと利用されるようになる	急速に利用されるようになる
流行り廃りが少ない	流行り廃りが激しい
熟練者から利用されやすい	初心者から利用されやすい

1.2.3. 本研究で対象とする情報

世の中の様々な情報に対して、長期的に利用されている情報を取得することの有効性が考えられるが、本研究では、特に以下のような情報を取り上げて、研究を行う。

- Web ページ
- 検索キーワード
- ファイル

図 1.1: 対象とする情報

本研究では、以下 3 点の特徴を持った情報として、これら 3 つの情報を取り上げた。

(1) 膨大な量の情報に手軽にアクセスすることができる

(2) 長期的に利用できる情報にアクセスすることが困難である

(3) 利用履歴など、情報の利用のされ方に関するデータを入手できる

以上の特徴を持った情報を対象とした理由を述べる。そもそもアクセス可能な情報量が少ない場合、所望の情報にアクセスすることがあまり困難ではないため、(1) の特徴を選んだ。また、本研究では、長期的に利用できる情報へのアクセスが困難になってきているという問題意識があるため、(2) の特徴を選んだ。さらに、情報の利用履歴など、情報の利用のされ方に関するデータがないと、長期的に利用されているかどうか判定できないため、(3) の特徴を選んだ。

また、これらの情報の中から長期的に利用されるものを見つけるためのシステムとして、以下の図 1.2 のように、3 つのシステムを提案し、設計・実装した。これらのシステムについての詳細および Web ページ、検索キーワード、ファイルを選んだより詳細な理由に関しては、第 4 章〜第 6 章で述べる。

1.2.4. 長期的な情報の具体例

本研究で述べている長期的な情報とは、具体的にどんな情報なのか、実際に具体例を述べる。

(1) Web ページ

長期的に利用する Web ページとして、以下のようなものなどが考えられる。

- まとめサイト

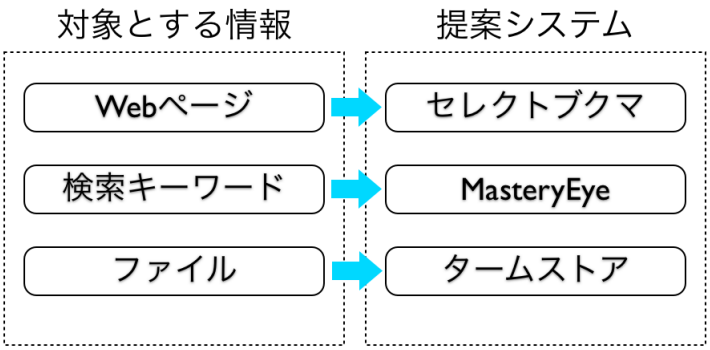

図 1.2: 対象とする情報と提案システム

- リファレンス的なサイト
- Web サービス

まとめサイトとは、技術や趣味などについてまとめて情報が記載されているサイトのことである。例えば、OS やソフトウェアのショートカットキーをまとめたサイト、美味しいレストランをまとめたサイト、論文の書き方などをまとめたサイトなどがあげられる。

リファレンス的なサイトとは、文章やプログラムを書く際に、リファレンス的に利用するサイトである。例えば、プログラムを書くときに利用する各プログラミング言語のリファレンスやプレゼン資料を作るときやデザインをするときに利用する画像や写真の素材集などがあげられる。

Web サービスは、例えば、検索サービス、動画共有サービス、レシピ共有サービスなどがあげられる。

(2) 検索キーワード

長期的に利用する検索キーワードとして、以下のようなものなどが考えられる。

- 各分野における重要キーワード
- ロングセラーとなっている商品名
- 長期間利用されている Web サービス名

各分野における重要キーワードとは、その分野について詳しく調べるときに必要となるキーワードである。例えば、学問、技術、趣味など様々な分野に対して重要キーワードが存在する。

ロングセラーとなっている商品名は、例えば食品や日用品などの各分野でロングセラーになっている商品名である。

長期間利用されている Web サービス名は、例えば写真や動画など様々な共有サービスや Q & A サービスなどのサービス名である。例えば、「YouTube[?]」や「Flickr[?]」などのキーワードは、検索キーワードとしては、長期的に利用されているだろう。

(3) ファイル

長期的に利用するファイルとして、以下のようなものなどが考えられる。

- テンプレートファイル
- リファレンス的なファイル
- 画像などの素材

テンプレートファイルとは、仕事などで書類などを作成する際に、雛形となるファイルである。例えば、Word、Excel、PowerPoint などの雛形ファイルがあげられる。何かの申請をするときや何かフォーマットの決まった書類などを作成するときに利用されることが多い。

リファレンス的なファイルとは、何かファイルを作成するときに、何度も参照するファイルのことである。例えば、文章を書く参照する以前書いた文章ファイルやデータファイルなどがあげられる。

画像などの素材とは、例えばよくプレゼン資料や Web サービスに登録するアイコンなどの素材である。こういった素材の中には長期的に利用されるものも多いだろう。

1.2.5. 長期的な期間について

本研究では、長期的に利用されている情報という表現をしているが、長期的とはどの程度の期間なのかについて述べる。一般的に、長期的とは絶対的に期間が決められているものではなく、どちらかというと相対的な表現として用いられる場合が多い。そのため、どういった情報について述べるかによって、期間が異なってくる。

そのため、1ヶ月以内しか利用されないものが多いタイプの情報ならば、3ヶ月でも長期的となるし、逆に半年ぐらい利用することが当たり前のタイプの情報ならば、半年利用していても長期的とはならない。そこで、本研究では、以下のようの平均値と標準偏差を用いて、長期的な情報と短期的な情報を示す。

長期的な情報：長期度 $>$ 長期度の平均値 $+$ 長期度の標準偏差
短期的な情報：長期度 $<$ 長期度の平均値 $-$ 長期度の標準偏差
※ここで長期度の平均値と標準偏差は、各種の情報ごとに計算する
例：Web ページ、検索キーワード、ファイルごと

以下の表 1.2 に、各情報についての具体的な期間を示す。

表 1.2: 長期的と短期的の期間

情報の種類	長期的	短期的
Web ページ	XXXX	XXXX
検索キーワード	XXXX	XXXX
ファイル	XXXX	XXXX

1.3. 長期的検索の提案

以上を踏まえて、本研究では、長期的な情報を検索するシステムを提案する。これは、検索者が今後も長期的に利用できる情報を見つけることができるようにすることを目指すものである。そのために、これまでに長期的に利用されてきた情報に特化して検索可能なシステムを実装する。

長期的に利用されてきた情報に特化して検索するために、まず、長期度の計算手法を提案する。そして、長期度の高い順に順位付けを行い、結果を提示する。