

コンテンツ検索エンジン ZDEA（ゼディア）の開発 ～ アクセス・ログによるコンテンツ評価アルゴリズムの実装～

開発代表者：山岡 幸作

1. 背景

高速なプロセッサ、大容量のハードディスク、定額制ブロードバンド回線の普及などによりインターネットには、未だかつてないほど多くの画像、音声、動画、マクロメディア・フラッシュなどのバイナリーコンテンツが公開されている。

そして、数百億ものウェブページから目的のウェブページを瞬時に抽出できる検索エンジンを利用できるようになった今日でさえも、目的のバイナリーコンテンツを探すという行為は、十年前と同じようにワールドワイドウェブを右往左往することしかできなかった。

2. 目的

このような問題を解決するべく、今までにない評価アルゴリズムを備えた検索エンジンが必要であると考えた。この新しい検索エンジンを開発することこそが本プロジェクト実施の目的である。

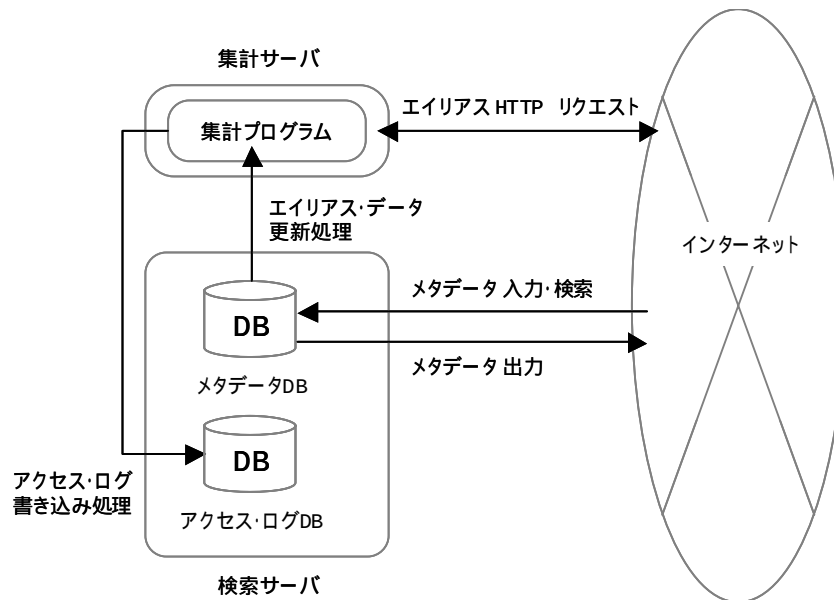
現在、ワールドワイドウェブにおいて検索するという行為は主に検索エンジンを利用する。検索エンジンは、クローラープログラムによるロボット型検索エンジンと専門家(サーファ－)によるディレクトリ型検索エンジンの二種類が有力である。どちらも十年前から存在している。

ロボット型検索エンジンはウェブページを収集し、文章(テキスト)の解析、被リンク数だけでなく、ウェブページのリンク構造による重みをもとにウェブページの評価する。ディレクトリ型検索エンジンは専門家が厳選したウェブサイトをディレクトリに分類する。

現在の有力な二種類の検索エンジンではバイナリーコンテンツの検索は困難である。本プロジェクトにおいてはリンクの構造に依存せず、かつ専門家よりも客観的な評価として、コンテンツごとのアクセス・ログの統計解析をもとにバイナリーコンテンツを評価するアルゴリズムを考案し、この評価アルゴリズムを実装するコンテンツ検索エンジンの開発を試みた。

3. 開発の内容

本プロジェクトの目的を証明するためには、最低 2 台のサーバが必要であり、この 2 台のサーバの開発を本プロジェクトの目標とした。2 台は以下の構成とした。



集計サーバ：高速ウェブサーバ

多数のコンテンツのアクセス・ログを取得するために、サーバには多量の HTTP リクエストによる負荷が発生する。本プロジェクト専用の高速ウェブサーバを開発した。

検索サーバ：コンテンツのメタデータとアクセス・ログを記録する DB サーバ

検索サーバは、コンテンツのメタデータの入出力インターフェイス、ログを記録する DB サーバとした。

コンテンツ・クリエイターが本プロジェクトの機能を利用するためには、HTML ファイルのコンテンツの URL を、エイリアス（本プロジェクトの URL）に書き換える。



4. 従来の技術（または機能）との相違

現時点では登録コンテンツ数が限られているために比較が困難であるが、今後は、コンテンツ登録の難易度など、それぞれの技術、サービスの得手不得手をより明確にしたいと考えている。以下は、既存のコンテンツ検索技術と本プロジェクトの比較である。コンテンツを検索する基準として「単位」「評価基準の客観性」「規模拡張性」の3点が重要であると考えている。

- | | |
|-----------|----------------------------|
| 1. 単位 | 単位が細かいほど検索精度が高まる。 |
| 2. 評価の客観性 | 評価アルゴリズムの客観性が高いほど検索しやすくなる。 |
| 3. 規模拡張性 | 検索対象が多いほど、ディスク容量や帯域を圧迫する。 |

サービス名	単位	評価の客観性	規模拡張性	総評
ZDEA	○ コンテンツ	○ アクセス解析	○ 特に制限はない	最適
写真サイト Flickr	○ コンテンツ	× 登録日時	○ 特に制限はない	○ 適している
配信サイト Apple iTunes, USEN Gyao	○ コンテンツ	○ アクセス解析	× 登録制限あり	○ 適している
Google, Yahoo! Search Technology	ウェブページ	○ 被リンク数、重み	10億枚の画像	質（精度）より量
ディレクトリ型検索サイト Yahoo! 登録サイト	× ウェブサイト	× ディレクトリ名	× 95%のサイトが登録拒否	× 不向き

5. 期待される効果

現時点では登録コンテンツ数が限られているため、コンテンツ・クリエイター（インターネットで自身のコンテンツ公開している創作者）への利点を本プロジェクトの特徴とする。

- ・コンテンツごとに RDF 形式のメタデータを入出力することが可能。
- ・コンテンツごとにアクセス・ログを取得することが可能。
- ・アクセスログを利用した検索を行うことが可能。

6. 普及（または活用）の見通し

本プロジェクトは今後も開発を続行する。プロジェクト実施期間においては、ユーザ・インターフェイスの難易度やサービスの難解さからクリエイターへの認知が難航した。今後は以下に重点を置く。

- ・インターフェイスの改善
- ・クリエイターへの認知活動
- ・法人営業

7. 開発者名

開発代表者：山岡 幸作（有限会社ハルジオン）

共同開発者：山澤 宏鑑（有限会社ハルジオン）

8. （参考）開発者 URL

ZDEA.com

<http://www.zdea.com/>

以上