

【DRニュース・011】：検索エンジンの仕組みと比較～IT 業界の90：9：1の法則

2016年03月14日発信

インターネットの情報検索で、日常、何気なく^{なにげなく}使っている「検索エンジン」を探索して見ました。
・・・ 「検索エンジン」開発の歴史や最適化対策など、いろいろな戦略が見えてきます。

1. 検索エンジンの役割と仕組み

現在、検索エンジンとしては、Google、Yahoo!、Bing(Msn)、Baidu(百度)・・・などが有名です。

まず、**検索エンジンの役割**^{やくわり}ですが、インターネットユーザーの大半が、関心のある事柄について何かしらのキーワードを打ち込み、検索を行い、目的のページ（1つとは限らない）に到達^{とうたつ}して、満足^{まんぞく}のいく情報を得ようと試みます。

日々行う検索に対して満足する結果が得られていれば、ユーザーはその検索エンジンを引き続き利用してくれますし、その結果としてより多くの検索が生まれ、広告のクリックが増え、最終的には検索サービス提供会社の収^{しゅう}益^{えき}につながります。

逆に検索結果の質が低く、満足^{まんぞく}のいくWEBページにいつまでたっても到達できないようなものであれば、ユーザーは別の検索サービスを使おうとするか、検索ではなく別の手段で情報を探しにいてしまうかもしれません。結果としては検索サービス提供側の収^{しゅう}益^{えき}が下がります。

したがって、**検索エンジンの役割**^{やくわり}としては、日々送られてくる検索に対し、
常に最善となるような検索結果をユーザーに提示することが必要となります。

インターネット上に公開されているデータを収集してインデックスを生成し、ユーザーの検索を受けてランキング付けが行われ、検索結果が表示されるまでの一連の流れは、次のようなものです。

- ① まず検索エンジンの**クローラー**（ロボット）がWEB上に存在するファイルを収集します。
- ② 収集したデータは、インデクサと呼ばれるプログラムによって検索エンジンが後々処理しやすい形になるようにデータを変換する処理が行われます。
- ③ インデクサによって処理されたデータを検索データベース（**インデックス**）に格納します。
- ④ ユーザーから検索が送られると、**形態素解析**^{けいたいそかいせき}により検索語を分解、その内容を把握します。
- ⑤ **ランキングアルゴリズム**に基づきインデックスされているWEBページのスコアリングを行います。
- ⑥ 点数の高い順にランキングを付け、最終的な調整がなされて検索結果に表示されます。

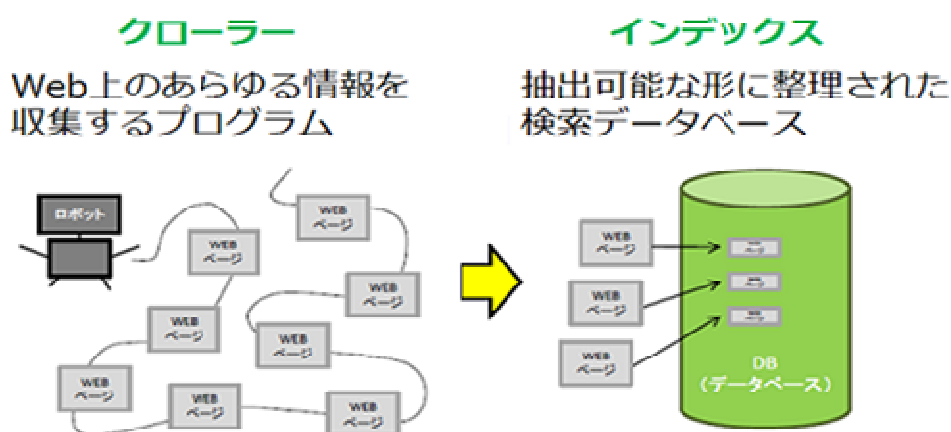
さらに、最善の検索結果を抽出^{さいてきか}するため、検索エンジンの最適化（SEO対策）が必要となります。

(1) クローラーとは

クローラー (Crawler) とは、Google などの**ロボット型検索エンジン**がWEB上のファイル (HTML 文書だけでなく、画像・PDF まで含む全般) を収集するためのプログラムのことです。

WEB上には、クローラーと呼ばれる検索エンジンのロボットが、常に巡回しています。

(本好きな人が世界中の全ての本を集めようとするように、WEB上の全てのページを集めようとします)



このクローラーが、あなたのサイトに回って来たとき、検索エンジンが初めてサイトの存在を知るため、あなたがサイトやページを作成したら、まず初めに、**クローラーに発見してもらうことが、必要！**

・・・ **SEO対策【Search Engine Optimization: 検索エンジン最適化】の第一歩となります。**

【クロールの最適化】

・・・具体的には、

- リンクが自然と集まる良質な (SEOコンテンツ) 作成
- XMLサイトマップの送信 ・・・などを行わなければなりません。

① リンク作成の対策

クローラーは、一つ一つのページに貼られている**内部リンク**をたどって、他のページに移動するため、ナチュラルなリンクが自然と増えるような良質なコンテンツを作ると、クローラーの巡回経路 (=外部からの被リンク) が増えるため巡回頻度があがります。したがって、ウェブ上の他の関係あるサイトに自分のコンテンツについて通知し、リンクを求めたり、自己が運営する既存のサイトから適切なリンクを貼ったりすることが対策として行われる。

② サイトマップの対策

サイトマップとは、サイトのウェブページのリストを指定して、Google や他の検索エンジンにサイトのコンテンツの構成を伝えるファイルです。Googlebot などの検索エンジンのウェブクローラーは、このファイルを読み込んで、より高度なクロールを行います。

クローラーによって収集されたデータがインデックス化され、巨大な検索データベースが作成されます。「ロボット」「スパイダー」などとも呼ばれています。

Google が持つクローラーには用途に応じて複数のクローラーがあります。

例えば通常の WEB 検索には「**Googlebot**」、画像検索には「**Googlebot-Image**」、モバイル検索には「**Googlebot-Mobile**」といったように、様々なクローラーが存在します。

(2) インデックス

クローラーが収集した WEB ページのデータを検索エンジンデータベースに（処理しやすいように整理された状態で）格納されることを「インデックス化」「インデクシング」などといいます。日本の SEO では常用語として「インデックス化される＝インデックスされる」と表現しています。

実際には検索エンジンクローラーが WEB ページのデータを収集した後にそのままの形でインデックス化されるわけではなく、一旦「インデкса」と呼ばれるプログラムによって検索エンジンにとって処理しやすいデータに変換する中間処理が行われ、その変換されたデータがインデックス化されます。

この時に、重要なのは、ページ自体を検索エンジンから見て読みやすいように整えておくことです。なぜなら、検索エンジンはインデックスされた情報をもとに後述する **ランキング** を行うからです。

文章が読みやすく書かれている本の方が、内容を理解しやすいのと同じで、適切に構成されているページの方が読みやすく、SEO 的に正当に評価されやすくなる。そのための施策をインデックス最適化といいます。

【インデックスの最適化】

Google のランキングエンジンがページの内容を正確に理解するように、ページの HTML 構造を最適化すること。または、検索結果に載らなくとも良いページはインデックスされないようにすること。

このためにまず重要なのは、WEB ライティングの基本を抑えて、**検索エンジンから見て読みやすいページ**を作ることです。さらに、noindex を有効活用して、**インデックスされるべきページとインデックスされなくとも良いページを区別することも重要となります。**

【コンテンツの最適化】 (最重要！)

Google から検索ユーザーからも感謝され、誰もが検索上位に表示されるべきと思うぐらい有益なページを作ること。そもそも、良質なコンテンツ（魅力的なコンテンツは自然に認知が広がります）がなければ、どれだけクローラー最適化やインデックス最適化に力を注いでも意味はありません。

なぜなら、クローラー最適化とインデックス最適化は、ページを本来あるべき順位よりも上に表示させることなく、**そのコンテンツの価値を適切に評価してもらうためのものだからです。**

(3) 形態素解析

形態素解析（morphological analysis）とは、検索エンジンにも用いられている自然言語処理の手法の一つで、ある文章・フレーズを「意味を持つ最小限の単位（＝単語）」に分解し、文章やフレーズの内容を判断するために用いられます。

具体例 例えば下記のような文章があったとします。

「SEOは、検索エンジンで上位表示させるための手法です」これを形態素解析で単語分解すると、
SEO / は / 検索 / エンジン / で / 上位 / 表示 / さ / せる / ため / の / 手法 / です /

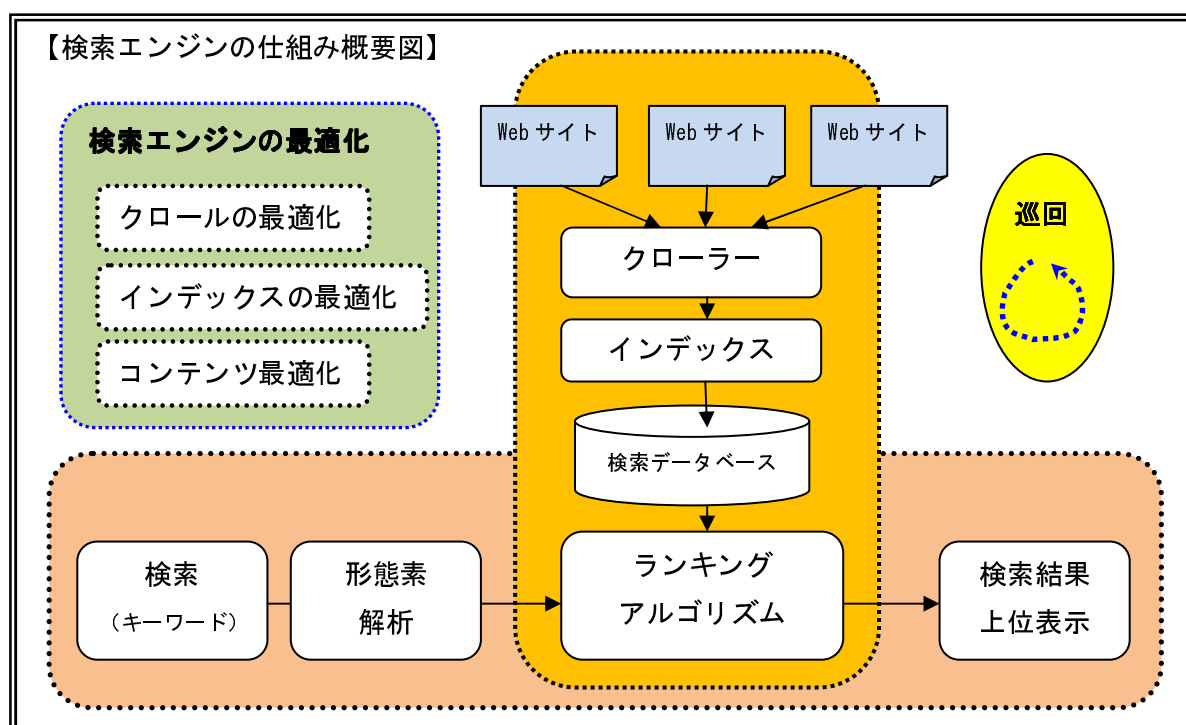
このように、文章を一つずつ品詞分解（名詞・動詞・助詞～）して文章がどのような単語で構成されていて、どのような意味を持つかを判断します。このような処理を行って、WEB ページがどのようなテーマを含んでいるかをキーワード毎の出現頻度や突出度などから判断して推測しています。

形態素解析の検索への応用例

検索エンジンは検索が行われる度に、その検索されたキーワードをそのまま用いるのではなく、形態素解析によって検索キーワードを分解し、不必要な言葉を省いて検索結果を抽出することがあります。これは、無駄な単語（検索結果に影響しないであろう単語）の処理を省略することで、無駄なデータ処理を軽減することが目的です。

ここで処理を省かれる単語のことを「ストップワード」と呼びます。

例えば、記号はストップワードになりますし、日本語では「で」や「が」などの助詞が省かれることがあります。



(4) ランキング・アルゴリズム

アルゴリズム(algorithm)は辞書をひくと「算法(さんぽう)」と訳されますが、日常的な言葉で言えば「計算方法」です。少し厳密^{げんみつ}に言えば、「ある特定の目的をより効率的に達成するために定式化された処理手順」で、コンピュータ用語として使われるのが一般的な「アルゴリズム」の用いられ方です。

SEO 対策で「アルゴリズム」と言えば、Google をはじめとする検索エンジンが検索結果で WEB ページにランキング付けをするための計算に用いられるランキングアルゴリズムを指すことがほとんどです。

このランキングアルゴリズムは、前段落^{まえだんらく}のアルゴリズムの解説の中の「特定の目的」を「検索されたキーワードに対して WEB ページのランキングを付けること」とすると、ランキング付けを適切に、かつ迅速^{じんそく}に処理するために検索エンジン側のプログラムの中に組み込まれた非常に複雑な計算手順です。

例えば Google では細かなアルゴリズム調整から大規模なアルゴリズムアップデートまで含め、多い時には年間で数 100 回に渡る変更が行われ、かつその具体的な内容についてはほとんどが公表されないため、多少のデータ検証を行ったところでその分析・解明を正確に行うことは不可能と言っても過言ではありません。

【主な Google のアルゴリズム概要】

- ① ドメインに関する項目；取得からの機関、情報表示の有無、ドメインの地域・種類など
- ② サーバーに関する項目；場所（日本／アメリカ）、信頼性
- ③ URLに関する項目；構造、見出しタグの使い方、外部 CSS や JavaScript の有無活用
- ④ コンテンツに関する項目；サイト内の使用言語の一致、リンクと文章の比率、文章の量、市場の専用性／ニッチ性、画像の ALT タグなどの設定、オリジナルコンテンツの有無など
- ⑤ 内部リンクに関する項目；ページ内の内部リンクの数、密度
- ⑥ サイト全体に関する項目；サイトの更新頻度^{こうしんひんど}、サイト内のページ数、インデックスされてからの時間、XML サイトマップの使用、お問い合わせやプライバシーポリシーの有無、ブログ？
- ⑦ ページに関する項目；タグの使用の有無や適切さ、ページ公開からの時間、ページの新鮮さ^{ひんぱん}（頻繁に追記）、コンテンツの読みやすさ、ページの表示速度・種類、ページの人気度
- ⑧ SEO キーワードに関する項目；タイトル内キーワードが使われているか？キーワードの有無や適切さ、キーワードの位置、ファイルの名前でのキーワード使用の有無、自然さ、適切さ
- ⑨ 外部リンクに関する項目；ドメイン当たりの外部リンク数、サイトやページの質、リンク先にスパムサイトなどが含まれて要るか？、リンク先が 404 になっていないか？、画像引用の有無
- ⑩ 被リンクに関する項目；リンク先サイトとの関係性・質、リンクの多様性、オーソリティ
- ⑪ 訪問者に関する項目；訪問数、訪問者のデモグラフィック情報^{りだつりつ}、離脱率、滞在時間などの傾向、訪問者の増減トレンド、検索結果内での広告のクリック率
- ⑫ ペナルティ項目；キーワードの詰め込みの有無、リンク購入・販売の有無、過去のスパム履歴^{りれき}、クローキングの有無、隠しテキストの有無、ペナルティの履歴、攻撃の履歴など

(5) ページランク

ページランク (Pagerank) とは、「リンクがより多く集まっている WEB ページはより重要である」という考え方に基き Google が開発した、WEB ページの重要度を示す一つの指標です。日本語では「参照重要度」などと表現されることもあります。

Google はリンクを「支持投票」として捉えます。多くの支持を受けるサイトほど人気があり、重要なものであるという判断基準があり、それを示す数値的な指標がページランクです。

2. 検索エンジンの最適化 (SEOとは)

(1) SEO対策の必要性

ユーザーが検索エンジンを使って調べたい項目を検索した時に、検索エンジンの順位に上位表示された方が当然ながらユーザーの目につきやすく WEB サイトの訪問者数（ページビュー）も増えます。企業などの商品販売ページでは、SEO 対策によって上位表示されることにより、莫大な収益を上げる例が少なくありません。

現在も [SEO 対策専門業者も多数存在しています](#)し、
購買意欲の高い検索キーワードで上位表示されることは大きなメリットがあります。

(2) SEO対策 具体的な進め方

SEO 対策は、「内部 SEO 対策」「外部 SEO 対策」の両輪からなります。できれば、この両輪を Google などの検索エンジンガイドラインに沿って実行すること（ホワイトハット SEO）が適切です。

① 内部 SEO 対策

- サイト名・記事タイトル・見出しタグ・画像・キーワードなどを適切に設定する
- サイト内部を閲覧者にわかりやすく案内する為リンクしたりしていく
- サイト内部に閲覧者が便利と思うコンテンツを増やしていく
- ・・・検索エンジンに良いサイトと認識してもらいやすくサイト内部を調整していきます。

② 外部 SEO 対策

- 外部サイトからのバックリンク数を増やす
- 外部サービス等からサイト URL を共有していく
- ・・・サイトで扱うテーマが同じで、良質なサイトからの被リンクは高評価されます。
(被リンクとは運営サイトが他 URL からリンクされている状況です)

- ・・・「[SEO チェキ](#)」は、WEB サイト URL を入力して、指定のキーワードをいれることにより簡単に検索エンジン順位を調べたり、ドメインエイジを調べたりできる[無料ツール](#)です。

3. 検索エンジンの比較

(1) 検索エンジンのシェア

日本では、世界でも珍しい？Yahoo! JAPAN が優勢です⇒**なぜ、Yahoo! JAPAN が強いのか？**

理由1：1996年設立のYahoo! JAPANに比べGoogleは1998年設立で市場参入が遅れた。

理由2：Yahoo! JAPANは日本独自のサービスを展開、世界共通と一線を画しています。

日本のシェア (2015年5月現在)	
Yahoo!	60%
Google	30%
Bing(Msn)	5%
その他	5%

世界のシェア (2015年5月現在)	
Google	65%
Baidu(百度)	8%
Yahoo!	5%
その他	22%

① 検索エンジンの現状

- 検索連動型広告が大きなビジネスになることから、「Yahoo!」がYST、「MSN」がLive Searchを開発。「Google」キラーと呼ばれる新興検索エンジンも登場しましたが、
..... **ふたたび淘汰の時代へと向かっています。**
- 2009年6月にマイクロソフトが次世代の検索エンジン「Bing」を投入。米Yahoo!とマイクロソフトの提携により、**米Yahoo!もYSTからBingに変更されること**となりました。
- しかし、**Yahoo!ジャパン**は、2010年7月に**Googleの導入**を発表し、日米による「ネジレ現象」が起こることになりました。
- しかし、今後はグーグルもヤフーに検索広告やアルゴリズムサーチ、イメージサーチを提供すると発表したんです。契約の期限は2018年までで、**グーグルとマイクロソフトの検索結果が表示される米ヤフーが、今後最強の検索エンジンになったりするのかな？**

② 検索エンジン開発の経緯 (2004年5月頃のニュース資料から抜粋)

- **なぜ米Yahoo!はYSTを開発したか**
米Yahoo!は2000年以降、今や世界最大の検索エンジン会社であるGoogleと提携をして検索サービスを提供していました。それより前はaltavistaと提携をしていました。
Yahoo!は当時、検索サービスをあくまでポータルが提供する単なる1つのサービスとしてしか考えていなかったため、他社と提携してサービス提供を実現していたのです。いしかえれば、検索そのものはビジネスになる、収益の得られるサービスとは捉えていなかったのです。
- **しかし状況は変わります。**Overtureの検索連動型広告での成功、そして検索エンジンを利用するユーザーの増加。検索エンジンがWebナビゲーションの中核をなし、Webサイトを訪問する際はまず検索エンジンというオンライン行動が顕著になりました。**つまり、検索エンジンがいわばWebサイト訪問のための”入り口”としての役割を果たすようになったわけです。**

(2) 主な検索エンジンの種類

検索エンジンには、Web 検索・カテゴリ検索・画像検索などで、いろいろな種類があります。

	Web 検索	カテゴリ検索	画像検索	その他
	Google／(YST)	独自	独自	動画・地図・商品
	日本でNo. 1 検索エンジン、※1Web 検索は JapanYahoo は Google に変更した			
	Google	DMOZ	独自	動画・ニュース・地図
	世界最大の検索エンジン、Web 検索の制度は定評有り、多くのサイトで利用			
	Bing	無し	独自	動画・ニュース・地図
	2009 年 6 月に検索エンジンを LiveSerch から Bing に変更した			
	Baidu	無し	独自	動画
	中国でNo. 1 検索エンジン、2008 年 1 月から日本で正式にサービス開始です			
	YetiBot	無し	有り	動画・クチコミ・テーマ
	韓国でNo. 1 検索エンジン、2009 年 6 月から日本でのサービスを再開です			

① ディレクトリ型（カテゴリ検索）サーチエンジン

ディレクトリ型（カテゴリ型）サーチエンジンは、登録申請のあったサイトを手動でデータベースに登録していきます。

登録する際に申請者がカテゴリやキーワード、紹介文などを入力し、担当者が内容の確認を行うため、ユーザーが検索した際にサイトの内容が適切でわかりやすく表示されます。

代表的なサイトは「Yahoo!」や「All About」ですが登録されるには審査があるため、簡単には登録されません。ただ登録されると Yahoo! や Google などのWEB 検索のSEO に大きく貢献します。

② ロボット型（ページ・WEB 検索）サーチエンジン

WEB 検索に使われるロボット型サーチエンジンは Yahoo! の「YST」と「Google」を使用するグループと独自のエンジンを持つ「msn」や「百度」などがあります。

ポイントはユーザーが入力したキーワードに対し、いかに良い品質の検索結果および順位を表示するかということになります。検索結果の向上やSEO 防止対策のためにアルゴリズムの変更がよく行われます。

4. 奇妙な比率や経験法則

テクノロジー業界につきまとう奇妙な比率「90 : 9 : 1 の法則」とは？

2016 年 02 月 06 日のニュース記事から抜粋

(1) インタネット上の「1%の法則」

インターネット上のコミュニティには、参加者よりも閲覧するだけの人 (ROM)の方がはるかに多いという「1%の法則」があると言われています。

これは、書き込を行う人は全体の 1%しかおらず、後は閲覧者であるという状況から来たものです。

(2) テクノロジー業界の「90 : 9 : 1 の法則」

これとは別に、テクノロジー業界においては「90 : 9 : 1」という別の興味深い比率にまつわる法則があると指摘されています。

OS やブラウザ、検索エンジンなどの IT サービスにおいて、市場に参入する企業のシェアが 90 : 9 : 1 の比率になるという「90 : 9 : 1 の法則」がささやかれています。

これは、最も人気を^{かくとく}獲得し多くの人の支持を得たサービスが市場の 9 割を独占し、2 番手が 9%を、3 番手でさえも 1%を確保するのがやっとなで、他のサービスは、数字上は存在しないに等しいという、じやくにくきょうしよく 弱肉強食の世界を指しています。

例えば、モバイル OS の市場シェアを例に挙げると、

Google の Android が約 83%、

Apple の iOS が約 14%、

Microsoft の Windows Phone (Windows 10 Mobile を含む) が約 2.6%で、

4 番手以下は 1%を確保できていないという状況で、おおむね「90 : 9 : 1 の法則」が成立します。

(3) パイレーツの法則「80 : 20」

経済において、全体の数値の大部分は、全体を構成するうちの一部の要素が生み出しているという理論。80:20の法則、ばらつきの法則とも呼ばれる。

(法則と言うよりもいわゆる経験則の類で、自然現象や社会現象は決して平均的ではなく、ばらつきや^{かたより}偏りが存在し、それを集約すると一部が全体に大きな影響を持っていることが多いとされる)

- 商品の売上の 8 割は、全商品銘柄のうちの 2 割で生み出している。→ロングテール
- 仕事の成果の 8 割は、^{ついやした}費やした時間全体のうちの 2 割の時間で生み出している。→集中力
- 故障の 8 割は、全部品のうち 2 割に原因がある。
- 所得税の 8 割は、課税対象者の 2 割が担っている。
- プログラムの処理にかかる時間の 80%は、コード全体の 20%の部分が^{しめる}占める。

5. ウィキペディアが Google に対抗する広告なしの検索エンジンを開発

2016 年 02 月 15 日のニュース記事から抜粋

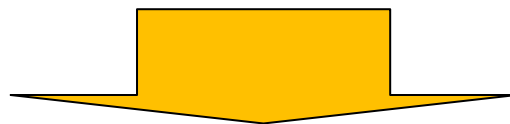
誰でも編集できるフリー^{ひゃっかつじてん}百科事典「Wikipedia」を運営している^{ウィキメディア}財団が、インターネット上で確実かつ信頼できる公開情報を見つけるための「Knowledge Engine by Wikipedia」を開発して、Google が圧倒的優位にある検索エンジンの世界へ打って出ようとしています。

ウィキメディア財団は、こうした Google などの商用検索エンジンがインターネットを「^{しはい}支配」して、インターネット上にある知識や情報を入手するための経路を、特許技術を使ってまで固めてしまっていると主張して、「Knowledge Engine by Wikipedia」の必要性を説いています。

「Knowledge Engine by Wikipedia」では、その^{情報の出所かどこなのか}がユーザーにもわかるように示され、^{情報の自由な流れ}のために^{経済界と完全に切り離すべく}広告も一切なし。

・・・「インターネット初の^{とうめい}『透明』な検索エンジンになる」とのことです。

ちなみに、Wikipedia では以前にも「Wikia Search」というオープンソースの検索エンジンを作ったことがあります。しかし、このプロジェクトはウィキメディアの設立者であるジミー・ウェールズ氏が求めていた結果が得られず、2009 年 3 月 31 日で打ち切り・^{へいき}閉鎖となった過去があります。



「検索エンジン」には、まだまだ、^{こんとん}混沌とした開発の背景があり、
ふたたび淘汰の時代へと向かっています

インターネット上の「検索エンジン」にも、時代背景やいろいろな技術、開発の経緯があります。

⇒今まで^{なにげなく}何気なく使っていましたが、そのロジックも日々更新がなされ、情報の発信源が何処なのかを常に意識して、正しい情報の集め方など「^{りょうしつ はんだん}良質な判断」をしていかなければなりません。