「Google の日本語解析と PSW 作成文章の有効性について」

―――――― グーグルは何を持って同一文章と判断するのか？ ――――――

あなたはミラーサイト（サイト全体がコピーされたもの）や、ホームページ内の文章の多くが他人のホームページの文章をコピペしたものであると、そのサイトがグーグルからペナルティーを受け、圏外に弾き飛ばされる可能性があるということを既にご存知で、そのことと PSW で作成された文章がかかわってくるのではないかとすごく心配されているかもしれませんね。

ひょっとしたら PSW の購入を検討するにあたって、その部分が引っかかっているかもしれません。

グーグルがどのようにして日本語を理解し文章全体を見分けるのか、あるいは同一文章と判断するのかについては、グーグルしか知らないわけで、いわばブラックボックスにあって誰もその答えを出すことはできません。

しかしながら**非常に重要なポイント**として、現段階でグーグルが日本語解析（認識や判断）する際に基本として取り入れているロジック（ルール）には、次の３つがあることがわかっています。

①形態素解析によって単語単位で分割処理すること。

②単語にはインデックスが付与されていること。（識別番号が付与されています。）

③Google エンジニアの論文には、「統計的に処理をしている。」と書いてあること。

少々小難しい内容になるのですが、わかりやすく説明しますと・・・

グーグルは文章を単語レベルで分割して、それぞれにインデックスを付与しています。
つまり単語に番号を付けているということです。

例えば「バスに乗る」という文章の場合は

バス　→　doc001
に　→　doc002
乗る　→　doc003

というように番号が付けられて、それがインデックスサーバーに保存されることになります。

そして例えば A と言う文章と B という文章があって、それぞれを比較する時にはインデックスサーバーに保

存されているインデックスを参照して

A 文章に　doc001 が 5 個　　doc002 が 15 個　　doc003 が 5 個
B 文章に　doc001 が 5 個　　doc002 が 15 個　　doc003 が 4 個

というような場合に同一文章として判断するだろうということです。

これが例えば

A 文章に　doc001 が 2 個　　doc002 が 10 個　　doc003 が 5 個
B 文章に　doc001 が 7 個　　doc002 が 6 個　　doc003 が 8 個

というような場合には同一文章とは判断されないと考えられます。

つまりグーグルが同一文章か否かを判断するのは、その文章に含まれている単語レベルでの同一性であり、かつその数が同一あるいは近接している場合に、同一文章として判断する可能性が非常に高いと考えられるわけです。

もっとわかりやすく言えば、単語レベルで違う文字が使われていたり、同一単語でも使用頻度が違っている場合には、全体として同一文章として判断されることはないだろうと言えます。

単語にそれぞれインデックスを付与しているのはそのような理由からだと推察できます。

また③の統計的に処理をしているということは、グーグルのこれまでの膨大な単語処理のデータによって、ひとつの単語に対しては、その次に来る単語や助詞や助動詞までをも推察する力を持っているということです。

このことは、グーグルで間違った検索をすると、

「もしかして○○ではありませんか？」

とか、間違った漢字で検索した場合にも、正しい漢字に直した検索結果を表示してくれますよね。

あなたも経験があると思いますが、それらは単語の近接、組み合わせを統計的に処理している証拠です。

グーグルのアルゴリズムには、この統計的な処理をする能力があるので、ワードサラダ的な文章（単語だけを繋げていて、日本語として成り立っていない文章）については、

「グーグルは間違いなく見抜くことができる！」

ということを意味しており、リライトされた文章が助詞や助動詞を含めておかしな文章になっている場合は、その文章は評価されないということを意味しています。

つまりでたらめな文章を含んでいるサイトの場合には、ペナルティーを食らって検索順位で圏外に弾き飛ばされることがあると容易に推測できるわけです。

もちろんそのようなでたらめな文章を含んでいるサイトから被リンクをもらっているメインサイトが上位表示されることもないだろうということも容易に推測できます。

上記のようなグーグルの日本語解析を理解したうえで弊社は PSW を開発しています。

PSW の基本的な考え方は、ひとつの原文を 10 個の文節に分け、それぞれを 10 個ずつリライトし、更にそれらをランダムに組み合わせて別の文章を作り上げるというものです。

従って、どこかにある文章をそのままコピーして持ってきて、それらを張り付けるような形で 1 ページを構成するというような類のツールではありません。

リライトするのはプロのライターで、固有名詞など別の単語に置き換えられないものを除いて、できる限り同じ単語にならないように気を付けてリライトしています。

『つまり単語レベルでの非同一性を目指しているわけです。』

別の言葉で言えば、大意（大まかな意味）は同じでも文字列を別のものに変更しているということです。

PSW はひとつの原文だけで上記のことを目指しているのですが、この原文が 100 個あり、それぞれをランダムに組み合わせて 1 ページ分の文章を作り上げてくれるのです。

当然、各ページにインデックスされる単語の数や種類は大幅に違ってくるわけです。

すなわち大意が同じということとコピー文章ということは、全く別の次元の話になるのです。

またグーグルの日本語解析の基本を鑑みても、PSW が作り上げる文章が同一文章だと判断されるのは非常に困難だと言えることがお分かりいただけると思います。

従って

「PSW により作成された文章がネット上にたくさん出現すると、
コピー文章と判断されペナルティーを受ける。」

という考えは当たらないと考えられます。

またグーグルの統計的な処理能力に対抗する意味で、PSW における原文のリライトを実施する時は、前後の単語の組み合わせについて非常に細かくチェックをしています。

つまり助詞や助動詞レベルで、日本語的におかしな文章にならないようにしているということです。

このことは出来上がった文章、ひいては出来上がったサイト全体が完璧な日本語文章として評価されることを目指しているのです。

評価されないことがわかっているワードサラダのサイトには決してならないようにしているわけです。

※上記の内容についてはプライベートな見解を含んでいることを予めご了承ください。

文：“PSW”事業部/田嶋　秀一

株式会社 T.P.L.Consulting

