Googleに隠れた秘密あり！KWIC方式の紹介文

視点(23) 2002年3月23日

1．Googleの躍進はランキングと紹介文にあり

Googleはリンクを利用したWeb用検索エンジンを1999年10月に正式オープンし、2000年8月から日本語Googleを提供しています。Googleはリンクを使うことでページ単位の検索エンジンに較べてコンテンツの情報量を飛躍的に増大させました。その結果、検索のランキング精度が向上し、多くの人から評価を受け、今日にいたっています。

リンクを使うとランキングが向上することがわかり、リンクを使った検索エンジンの開発が各所でなされるようになりました。リンクの利用といってもまだ模索段階であり、いろいろな方法を採用しています。ランキングは収集したデータによっても影響しますので、検索サイトのランキングが同じになることはありません。

検索デスクでは、2001年5月から検索エンジンの検索結果を評価しています。検索結果には、ランキング、紹介文、関連情報、キーに関連した情報、などがあり、各検索サイトごとに異なっています。検索デスクが行なう評価は出力したURLだけからであり、紹介文などを対象にしていません。検索力調査の評価では、1位goo(米Inktomi系)、2位Google、3位Lycos(米WiseNut系)が4位以下を大きく離しています。URLの評価はあまり優劣がつかないのですが、利用者はGoogleを高く評価します。

GoogleはオープンしたときからKWICK方式の紹介文を採用しています。それがGoogleの評判に貢献していますが、誰もそこに気づいていません。今回の視点では検索結果の表示内容について、私なりにまとめました。

2．検索結果の表示内容

表2-1　検索サイトの検索結果の表示内容（数字は順位）

検索サイト	タイトル	紹介文	URL	キャッシュ	関連ページ	サイト内頁	更新年月日	適合率	その他
goo	1	5	3	-	-	-	4	2	-
Google	1	2	3	6	7	8	5	-	4容量
Lycos	1	2	3	4	-	-	-	-	-
Alltheweb	1	2	4	-	-	3	-	-	5容量
AltaVista	1	2	3	-	5	6	-	-	4翻訳
Naver	1	2	3	4	5	6	-	-	-
Infoseek	1	3	4	-	-	6	5	2	-
TOCC	1	2	5	-	-	-	4	3	-

検索すると検索結果が表示されますが、その表示内容は各検索サイトごとに異なります。それを表2-1にまとめました。縦は検索サイト、横は表示内容を、そして数値は表示順を示します。例えば、gooは、タイトル、適合率、URL、更新年月日、紹介文、の順に表示します。表示内容の変更がなされてもデータが新しくなるまでに時間がかかるのと、サイト表示とページ表示とが

異なる場合があります。

ユーザーがコンテンツを見つけるのに重要な手がかりになるのは**タイトル**と**紹介文**です。これについては後で詳述します。**URL**はドメインの知識がある人に出所についての情報を与えます。**キャッシュ**はGoogleが始めたもので、KWIC方式の紹介文作成の副産物で、リンク切れの際に役立ちます。なお、LycosとNaverのプレビューはキャッシュと同じです。

関連ページは類似したページを表示します。**サイト内ページ**は同じ情報源のページをまとめて表示し、冗長さを防ぎます。**更新年月日**は古くからあったのですが、最近Googleも採用しています。**適合率**も古くからあり、適合率の多少で検索キーと検索結果との関連性を示します。その他にある**容量**はコンテンツの容量をKBで表示しますが、グラフィックや広告を含めるとさらに大きくなります。**翻訳**はページ全部を翻訳します。以下、タイトルを３節、抄録方式の紹介文を４節、KWIC方式の紹介文を５節で詳述します。

3. 検索結果のタイトル

Webの表示内容はHTML文で書きます。それを書くルールが厳密でないためにWebは急成長しました。しかし、それを収集して処理する検索エンジンにとっては逆に大きな負担になっています。Webのタイトルに関しては2種類あります。一つはブラウザのトップに示されるもので、タグのTITLE内に書きます。もう一つはブラウザの表示画面に示されるもので、真っ先に利用者が目にするものです。

画面に表示されるタイトルはタイトルを示すタグはなく、文字だけでなくグラフィックで示されたりします。タイトルは制作者が一番強調したいものですが、タイトル自体がないのもあります。従って、検索エンジンがタイトルを作成するには後者を使うことはできず、結局、内容を的確に表現しているとは限らない前者のTITLEタグを使用します。一般に、サイトのタイトルは短く、ページのタイトルは長くなる傾向があります。

タイトルの文字数を調べるために、いろいろなキーで検索してみました。わずか100サンプルですが、その平均は全角で13文字でした。Webと異なる論文の世界において、一つの論文はタイトル、抄録、本文、出典で構成します。タイトルは本文の60%位の内容を理解できるだけの情報量を含むように表現します。新聞なども同様です。それに較べるとWebのタイトルは短かすぎてコンテンツを理解するのには不足です。

タイトルの良し悪しは検索エンジン側ではなく制作者側にあります。TITLEタグは全文採用されますので、少なくとも画面の1行に収まる全角30字位のタイトルを書けば、検索エンジンのランクがあがり、利用者からのクリック率は高くなり、アクセスアップにつながります。

4. 抄録方式による紹介文の作成

紹介文はサイトやページの内容を100字位で説明するものです。検索対象を選択する手がかりになりますので、タイトルとても重要な表示項目です。内容を圧縮する手段として情報検索の世界では、タイトルを補佐するものとして抄録が使われました。その延長でWebの紹介文は抄録とみなされました。タイトルがその機能を発揮していない分、タイトルよりも重要な情報になっています。紹介文の作成法はここで取り上げる抄録方式と次節のKWIC方式の2種類があります。

抄録はメタタグのdescriptionに書くのですが、細部のページになるとdescriptionを記したページは少なくなります。実際に調べてみると、descriptionのあるのは約15%、すなわち、7ページのうち1ページしかありません。これは検索エンジンにとっては負荷になり、したがって抄録を作成しなければならなくなります。タイトルの場合と同様に、制作者側による抄録の不備を検索エンジン側が負うばかりか、逆に出来が悪いと非難されたりします。この点、抄録を管理

できるディレクトリは有利な立場にあります。

さて、抄録の作成ですが、多くの場合、HTML文からタグを除いてできる日本語の文字列を最初から約100字分を自動的に拾い出して作ります。調べてみればすぐ分かりますが、このルールではジャンプ先、記号、広告、スタイルなどを含むため、とても抄録といえないものを作成してしまいます。

最近、最初の部分をカットして文章部分を100字ピックアップするようなものも開発されています。この方法では文章になりますが、まだ抄録とは言えません。なお、各検索サイトごとの紹介文の作成法を表4-1に、その表示内容を付録Aにリストしました。

表4-1　抄録方式による紹介文の作成

検索サイト	紹介文の作成
goo	descriptionを使用、ないときは段落前の文章から100字
Google	descriptionを使わずにタグを除いた最初から100字、 ディレクトリがあればその抄録やカテゴリを付加
Lycos	descriptionを使わずにタグを除いた最初から100字
Alltheweb	descriptionを使用、ないときは最初の文章から80字
AltaVista	descriptionを使用、ないときはタグを除いた最初から80字
Naver	descriptionを使わずにタグとジャンプを除いた80字
Infoseek	descriptionを使用、ないときはタグを除いた最初から100字
TOCC	descriptionを使用、ないときはタグを除いた最初から100字

サイトやサイト内ページなどを表示する場合には、この抄録方式の紹介文が必要です。ある程度の時間やコストをかければ、現在よりもよりよい抄録を作成できるかも知れませんが、コンテンツの質などを考えると非現実的です。最近、Googleはディレクトリの中に入っているコンテンツはその抄録やカテゴリを一緒に付加しています。現状ではこれが一番よい方法かも知れません。

5．KWIC方式による紹介文の作成

1999年にオープンしたGoogleは従来とは異なる方法で抄録文を作っています。1999年9月26日の検索力調査には、「Ringringは.....各ページの紹介文を作成する代わりに、全文を保持してその位置情報からKWIC的な紹介文を作成しています。Googleも全文を保持して紹介文を作成しています。」とあります。これによると、2001年12月末にクローズしたkensaku.jp(Ringring)も同様なKWIC方式の紹介文を提供しており、この業績を評価されることなく休止しています。

情報検索の世界ではKWIC索引という索引法があります。KWICは"KeyWord In Context"の略号で、キーワード前後の文脈を一緒に表示する索引です。18年前になりますが、私もKWIC索引のプログラムを作り、画面に表示(表5-1)したことがあります。

表5-1　KWIC索引の例（"+"があればそれ以降の"="から開始し、前に戻る）

+E: A CITATION ANALY =	BIBLIOMETRICS IN INFORMATION SCIENC
A COMPARISON OF A	BIBLIOMETRIC APPROACH AND AN HISTOR
+RNALS IN COMMUNIC = A	BIBLIOMETRIC EVALUATION OF CORE JOU
+ICAL RESEARCH = A	BIBLIOMETRIC ANALYSIS OF PHARMACEUT
+LLABORATION: A REVI =	BIBLIOMETRIC STUDIES OF RESEARCH CO
+NDATIONS, METHODS A =	BIBLIOMETRICS - ITS THEORETICAL FOU

Web以前の情報検索の世界では、タイトルやキーワードしか処理できませんでしたが、Webの世界では本文の全文検索が可能になっています。KWIC方式の紹介文はKWIC索引の原理を応用して、本文の中にあるキーワード前後の文脈から作成します。この方法は検索キーに対応して作成しますので、必ず検索キーは存在し、抄録よりも情報を探しやすくなります。絞込み検索を多用するプロの利用者に人気が高いのは納得できます。この実現にはページの全文を蓄積する必要があり、それを2次利用したのがキャッシュやプレビューとみなせます。

KWIC方式の検索サイトを表5-2に、その表示内容を付録Bに示します。これ以外の検索サイトは抄録方式による紹介文を表示します。検索キーの順序を変えると、Googleは紹介文も替わる場合があります。ページ内に複数のキーワードがある場合、どのキーワードの前後を出力するかなど、まだ解決すべき課題があります。また、紹介文は抄録方式だけとかKWIC方式だけでは万能でなく、検索状況に応じて両者を使い分けて表示することが必要です。

表5-2　KWIC方式による紹介文の作成

検索サイト	KWIC方式の紹介文の作成
Google	検索キーを含む文章を100字
Lycos	検索キーを含む文章を110字
Naver	検索キーだけリスト、まだ処理は不完全？

6．検索精度の向上を求めて

Googleの隠れた秘密を明らかにしたのですが、理解できましたでしょうか。KWIC方式の紹介文を開発した人たちはもっと高く評価されてよいのではと思います。ランキングがよくなり探している情報が目の前に表示されても、タイトルや紹介文が的外れであれば見過ごしてしまいます。

これは検索サイトにとっても利用者にとっても不幸なことです。上手く検索するにはランキングとタイトル・紹介文との両方がよくならなければなりません。KWIC方式は原理が明確ですので実現は早いですが、むしろ抄録方式の実現は時間がかかりそうです。それにコンテンツ制作者が少し長めのTITLEタグをつけるなどして協力しなければ、検索精度の向上は望めないように思います。

付録A　抄録方式の表示内容　検索キー「抄録　コンテンツ」

■goo -- 抄録方式 -- 段落直前の文章から100字
検索の視点 #22: 第3世代Web検索エンジンについて （検索デスク）
98% v20010705　2002/02/14
ここでは、その進歩の足取りをたどり、現状の検索エンジンについて理解を深めたいと思います。ロボット系といわれる検索エンジンは、(1)Webページを収集し、(2)それを索引化してデータベースを構築し、(3)検索要求に...

■Google -- 抄録方式 -- 最初から100字、要約とカテゴリ
検索の視点 #22: 第3世代Web検索エンジンについて ...
検索デスク　　検索の視点 #22, ...
v20010705 - 14k - キャッシュ - 関連ページ
[他、内のページ]

■Lycos -- 抄録方式 -- 最初から100字
検索の視点 #22: 第3世代Web検索エンジンについて （検索デスク）
検索デスク　　検索の視点 #22 ホーム　サイトマップ　ヘルプ　. 検索 ページ サイト 逆リンク URL 新聞 海 外 情報 検索力 記事 視点 動向 掲載 履歴 目次　視点:10 第3世代Web検索エンジンについて 2001年7月5

日 ...
v20010705 [プレビュー]

■Alltheweb -- 抄録方式 -- 文章の始まるところから80字
検索の視点 #22: 第3世代Web検索エンジンについて （検索デスク...
Web検索エンジン開発の歴史はインターネットの開始から数えても・・・6～7年しか経っていません。この間にWebコンテンツが爆発的に増加したため...
v20010705 (13.7 kB)

■AltaVista -- 抄録方式 -- 最初から80字、#22がないので#21を使用
検索の視点 #21: サイト検索を考慮した検索力調査 （検索デスク）
. 検索デスク. . 検索の視点 #21. ホーム. . サイトマップ. . ヘルプ. . 検索 ページ サイト 逆リンク URL 新聞 海外 情報 検索力 記事...
v20010519 ? Translate

■Naver -- 抄録方式 -- 最初からジャンプを除いた80字
{2] 検索の視点 #22（検索デスク）
検索デスク 検索の視点 #22 ホーム サイトマップ ヘルプ_第3世代Web検索エンジンについて 2001年7月5日 1 急発展中のWeb検索エンジン Web検索...検索...
v20010705 [Preview] [類似文書検索]

■Infoseek -- 抄録方式 -- 最初から100字ですがHTML文のバグを処理
検索の視点 #22: 第3世代Web検索エンジンについて 検索デスク : 40%
A { font-size: 12pt; line-height: 115%; text-decoration: none; } A:hover { background-color: #FFCCCC; } .b0 { color: blue; font-size: 10pt; line-height: 120% } .b1 { color: blue; ...
v20010705 2002.02.14 更新 www.searchdesk.com内の検索結果を全て表示▼

■TOCC -- 抄録方式 -- 最初から100字
検索の視点 #22: 第3世代Web検索エンジンについて （検索デスク）
検索デスク 検索の視点 #22 ホーム サイトマップ ヘルプ . 検索 ページ サイト 逆リンク URL 新聞 海外 情報 検索力 記事 視点 動向 掲載 履歴 目次 視点:10 第3世代Web検索エンジンについて
98% 2002/02/14 v20010705

付録B KWICK方式の表示内容 検索キー「抄録 コンテンツ」

■Google -- KWICK方式 -- 検索キーを含む文章を100字
検索の視点 #22: 第3世代Web検索エンジンについて ...
... の作成は困難でした。例えば論文の場合、**コンテンツ**を読んでそれを短い**抄録**にしなければならず、**コンテンツ**は濃縮されました。データベースの作成は時間とコストのかかるものでした。 分野により属性が異なるため ...
v20010705 - 14k - キャッシュ - 関連ページ

■Lycos -- KWICK方式 -- 検索キーを含む文章を110字
検索の視点 #22: 第3世代Web検索エンジンについて （検索デスク）
... するディスクリプタなどを使いました。特に、**抄録**の作成は困難でした。例えば論文の場合、**コンテンツ**を読 んでそれを短い**抄録**にしなければならず、**コンテンツ**は濃縮されました。データベースの作成は時間とコストのか ...
v20010705 [プレビュー]

■Naver -- KWICK方式 -- 検索キーだけリスト、まだ処理は不完全？
検索の視点 #22（検索デスク）
...抄録...抄録...抄録...コンテンツ...コンテンツ... 3 第１世代はコンテンツ...コンテンツ...
v20010705 [Preview] [類似文書検索]