自然言語によるユーザフレンドリーな Web 検索

目木信太郎†

† 岡山県立大学情報工学部　〒 719–1197 岡山県総社市窪木 111
E-mail: †meki@cse.oka-pu.ac.jp

あらまし　インターネット上には非常に数多くの Web ページが存在するので、多くのユーザはサーチエンジンを利用することによって、必要な情報が書かれている Web ページの位置情報 (URL) を得ている。しかしながら多くの場合、ユーザはただ一つのキーワードしか入力しないので、サーチエンジンは膨大な数の URL を返し、それゆえ必要な情報を得ることができない場合も多い。そこで、キーワードの選択をユーザではなく、サーチエンジン側で行なえばよりよい検索結果が得られるのではないかと考えられる。すなわちユーザが必要とする情報を、キーワードではなく日本語や英語などのような自然言語で表現するのである。そして、サーチエンジンのフロントエンドプロセッサがそれをいくつかのキーワードに変換し、それらのキーワードを利用して Web ページの検索を行なうのである。本稿では例を用いてその有効性を示す。

キーワード　サーチエンジン, キーワード, 自然言語処理

Shintaro MEKI†

† Faculty of Computer Science and System Engineering, Okayama Prefectural University
Kuboki 111, Soja-shi, Okayama, 719–1197 Japan
E-mail: †meki@cse.oka-pu.ac.jp

1. はじめに

インターネットは現代の情報化社会の重要なインフラストラクチャとなった。インターネット上の情報は、多くの場合 Web サーバ上に存在することが多い。しかし、インターネット上には数多くの Web サーバが存在し、しかもそれぞれの Web サーバにも数多くの Web ページが存在する。ユーザはどの Web サーバに必要とする情報が存在し、その Web サーバのどの Web ページに必要とする情報が書かれているかという位置情報 (URL) を知っていないとインターネット上の必要な情報を利用することができない。そこで、多くのユーザはサーチエンジンを使って URL の検索を行なっている。この場合、サーチエンジンに与えるキーワードを適切に選ぶことは非常に重要である。しかしながら、ユーザはサーチエンジンを使う時、たった一つのキーワードしか入力しないことも多い。この場合、非常に多くの URL が検索結果として返され、その結果ユーザは本当に知りたい情報を手に入れることができない場合もある。ユーザがたった一つのキーワードしか入力しない理由の一つは、キーワードの選択にはある種のコツがあるからである。多くのユーザはそのコツを知らず、それゆえ最初に思いついたキーワードを一つだけ入力しているのである。

キーワードの選択に当たってはもう一つの方法が考えられる。すなわちユーザはキーワードの選択を行なわず、サーチエンジン側が代わりに行なうのである。ユーザは検索したいことを日本語や英語のような自然言語で記述し (以後キーセンテンスと呼ぶ)、サーチエンジンのフロントエンドプロセッサがそれをいくつかのキーワードに変換する。その後で、サーチエンジンはそれらの変換されたキーワードを使って Web ページの検索を行なうのである。多くの場合、ユーザはたった一つのキーワードを使う場合よりも、より本当に知りたい情報を手に入れることができると考えられる。

本稿の残りの部分は以下のように構成されている。次の章では、関連する事例について述べる。第 3 章では我々が提案する二つのタイプのサーチエンジンを提示する。第 4 章ではシステムの構築法について述べる。第 5 章では実行例によって提案する手法の有効性を示す。最後の章は本稿のまとめを行なう。

2. 関連研究

この章では、研究ではないのだが、関連する事例について述べる。

検索サイトの Goo [1] ではキーワードだけではなく、日本語の文章を入力しても検索ができるようになっている。具体的な検索アルゴリズムは書かれていないのでよく分からないのだが、実際に使ってみたところでは、不要な単語を除いて残った

単語をキーワードとし、検索を行なっているようである。我々が提案する手法の単純なアーキテクチャに相当すると考えられる。我々は、そのような単純な方法だけではなく、日本語の意味も考慮した手法を提案しているので、この点において Goo とは異なる。

3. アーキテクチャ

3.1 単純なアーキテクチャ

キーセンテンスをキーワードに変換するもっとも簡単な方法は、助詞などの不要な単語、言い換えるとストップワードを取り除くことである (図 1)。

この方法は、キーセンテンスの意味を考慮していないので、不適切なキーワードを選択してしまうかも知れない。その理由の一つとして、サーチエンジンはあること (もの) そのものを検索するのではなく、あること (もの) が記述された Web ページを検索するのであるということがあげられる。

例えば、低反発クッションを販売しているオンラインショップを検索しようとして、ユーザがキーセンテンスとして、「低反発クッションを販売するオンラインショップ」と入力した場合、おそらくうまく検索できない。この場合、サーチエンジンのフロントエンドプロセッサはキーワードとして、「低反発クッション」、「販売」、「オンラインショップ」をキーワードとして選ぶであろう。しかしながら、オンラインショップの Web ページに必ず「オンラインショップ」と書いてあるわけではないし、まして、オンラインショップは何かを販売する Web サイトなので、わざわざ「販売」と書いてあるわけもない。

しかし、低反発クッションを販売する Web ページにはそれらの価格と、購入する数を入力する欄があるはずだということを考慮して、「低反発クッション」、「価格」、「数量」をキーワードとして選べばおそらくうまくいくだろう。すなわち、キーセンテンスの意味も考慮してキーワードに変換するのである。この場合、キーセンテンスの意味を理解して、さらにそれが Web ページ上にはどのように書かれているかを知っておく必要がある。

3.2 改良されたアーキテクチャ

図 2 はキーセンテンスをキーワードに変換する時、キーセンテンスの意味も考慮する改良されたアーキテクチャの概念図である。サーチエンジンのフロントエンドプロセッサはキーセンテンスの意味を解析するために演繹データベースに対して質問を発する。演繹データベースは、「もし何かを販売する Web ページならば、そこにはその価格が書いてある」とか、「もし何かを販売する Web ページならば、そこには購入する数量を入力する欄がある」などといった演繹ルールを持っている。もちろん、インターネット上には非常に多くの種類の Web ページが存在するので、演繹データベースはそのすべての種類の Web ページをカバーする演繹ルールを持つことはできない。しかしながら、ユーザの検索履歴を使うことによって、多くの場合に対して演繹データベースを適応させることは可能であると考えられる。

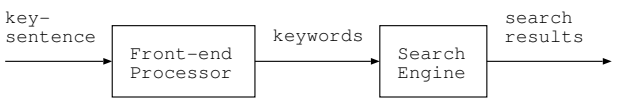

図 1 単純なアーキテクチャ

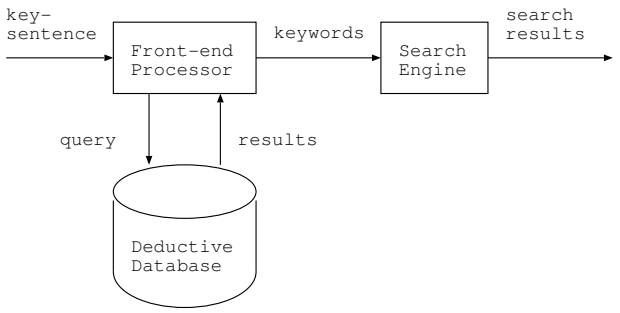

図 2 改良されたアーキテクチャ

4. システムの構築法

4.1 フロントエンドプロセッサ

この節ではフロントエンドプロセッサの構築法について述べる。

フロントエンドプロセッサは、入力されたキーセンテンスについて形態素解析を行い、助詞などのキーワードには成り得ない不要な単語を取り除く。残った単語をキーワードの候補とし、以下これらの単語をキーワード候補と呼ぶことにする。

フロントエンドプロセッサは、演繹データベースに対してキーワード候補の集合を質問として発する。そして、演繹データベースから返された質問結果を、キーワードの集合としてサーチエンジンに渡すようにする。

4.2 演繹データベース

この節では演繹データベースの構築法について述べる。

演繹データベースの構築法とは、演繹ルールの作り方と考えることができる。3.2 節で例として挙げた、「もし何かを販売する Web ページならば、そこにはその価格が書いてある」という規則の場合、三段論法を使ってもう少し論理的に書くと、「オンラインショップには何かを販売する Web ページがあり、何かを販売する Web ページにはその価格が書いてある」となる。

記号を使って書き直すと、

オンラインショップ → 何かを販売する Web ページ
何かを販売する Web ページ → 価格

と書くことができる。

同じようにして、「もし何かを販売する Web ページならば、そこには購入する数量を入力する欄がある」という規則の場合だと、

オンラインショップ → 何かを販売する Web ページ
何かを販売する Web ページ → 数量

となる。

また、何かを販売するところはオンラインショップに限らないが、オンラインショップは何かを販売するところに決まっているので、

{ オンラインショップ, 販売 } → { オンラインショップ }

という演繹ルールも考えられる。

このようにして演繹ルール、つまり演繹データベースを構築していくことができる。

この例から分かるように、すべての演繹ルールは、次のいずれかの形をしているので、矛盾は生じないと考えられる。

サイトのカテゴリを示す単語 →

そのサイトに存在する Web ページ

あるサイトに存在する Web ページ →

その Web ページに書かれていること

二つのキーワード候補 → いずれか一つのキーワード候補

ところで、演繹ルールを導き出す一般的な方法であるが、残念ながらそのような方法は存在しない。なぜなら、演繹データベースが持っている演繹ルールは、キーワードを決めるコツを演繹ルールで表現したものであり、キーワードを決めるコツを導き出す一般的な方法が存在しない以上、やはりそれに対応する演繹ルールを導き出す一般的な方法も存在しない。全ては、演繹データベースの作成者の技量にかかっているのである。

5. 実 行 例

3.1 節で、低反発クッションを販売するオンラインショップを検索する場合に「低反発クッション」、「販売」、「オンラインショップ」をキーワードに選んでもうまくいかないと述べた。確かに、最初にこのアイデアを思いついた時にはそうであったが、その後のサーチエンジン最適化 (SEO) [3] の普及にともなって、これらのキーワードで検索を行なってもうまく検索できるようになっている。それでは、キーセンテンスの意味を考慮しなくてもすべての場合にうまくいくのか、というとそれではうまくいかない場合も存在する。この章ではそのような場合の例を一つ提示する。

多くの場合、サーチエンジンは数十万の検索結果を返してくるが、ユーザはせいぜい上位 10 件の検索結果しか見ないので、ここでは上位 10 件の検索結果についてだけ考察を加える。

例えば、キヤノンの一眼レフデジカメ「EOS Kiss DIGITAL X」を販売するオンラインショップを検索する場合を考える。前々章で提案した単純なアーキテクチャによる検索では、キーワードとして「EOS Kiss DIGITAL X」、「販売」、「オンラインショップ」が選ばれることになる。これらの単語をキーワードとして、Google [2] を使って実験を行った。この場合図 3 のような検索結果が得られた。検索結果から直接または間接的に「EOS Kiss DIGITAL X」を販売するオンラインショップの Web ページに到達できたのは、上位 10 件の検索結果のうち 2 つだけであった。

一方、改良されたアーキテクチャによる検索では、キーワードとして「EOS Kiss DIGITAL X」、「価格」、「数量」が選ばれることになる。この場合図 4 のような検索結果が得られた。検索結果から直接または間接的に「EOS Kiss DIGITAL X」を販売するオンラインショップの Web ページに到達できたのは、上位 10 件の検索結果のうち全部で 4 つであった。

顕著な違いが出たとはやや言いがたいが、それでも単純なアーキテクチャの場合の 2 倍の Web ページを見つけることができた。このことから、キーセンテンスの意味を考慮した方がよい結果が得られる場合があるということがわかる。

ところで、読者の中には実行例の選択が不適切ではないかと思う人もいるかもしれない。確かに、すべてのオンラインショップの Web ページに「価格」と「数量」の両方が書かれているとは限らない。例えば「数量」は別の Web ページで入力するようになっているかも知れない。しかし、インターネット上には無数のオンラインショップが存在し、「価格」と「数量」が同じ Web ページに書かれている場合と、別々の Web ページに書かれている場合の割合は誰にもわからない。実験結果からうまくいく場合が少なからず存在することはわかるので、例えすべての場合を調べていなくても提案した手法にも有効性はあると考えられる。我々は何もすべてのオンラインショップが漏れなく検索できると主張しているのではないのである。

6. 結 論

本稿では、キーワードの代わりに自然言語を利用する 2 つのタイプのサーチエンジンのアーキテクチャについて述べた。改良されたアーキテクチャがうまく働くためにはキーセンテンスの意味を理解する必要がある。さらに、演繹データベースはそれらの意味がインターネットの Web ページにどのように表現されているかを知っておく必要がある。インターネット上には非常に多くの種類の Web ページが存在し、ユーザの要求もそれぞれのユーザによって異なる。それゆえ、演繹データベースは常識のようなものを扱うことになるので、提案した手法をすべての場合に適応させるのは不可能と考える読者もいるかも知れない。

しかしながらそれは正しくないと考えられる。なぜなら、改良されたアーキテクチャの場合、もし演繹データベースに質問とマッチする演繹ルールが見つからなければ、質問として与えられたキーワード候補の集合をそのまま質問結果として返すことができるからである。そして、そのような結果を受けとったフロントエンドプロセッサは、次善の策として単純なアーキテクチャと同じキーワードをサーチエンジンに渡すことができるからである。だから、4. 章で示したようなたった 4 つの演繹ルールしか持たないような演繹データベースでも理論上はうまく動作する。

また、演繹データベースが返す質問結果は、より良いキーワード候補の集合か、あるいは、質問として与えられたキー

（1） ツクモネットショップ
http://shop.tsukumo.co.jp/
EOS Kiss DIGITAL X を販売する Web ページへのリンクあり
（2） 翔泳社、オンラインショップ“ SEshop.com ”をオープン
http://ascii24.com/news/i/serv/ article/1999/07/01/603122-000.html
EOS Kiss DIGITAL X に関する記述無し。キャッシュには EOS Kiss DIGITAL X のレビュー記事の Web ページへのリンクあり
（3） トイザらス、おもちゃのオンライン販売サイトをオープン
http://ascii24.com/news/i/mrkt/article/ 2000/10/03/618567-000.html?geta
EOS Kiss DIGITAL X に関する記述無し。キャッシュにも記述無し
（4） デジイチ最新情報！ — キヤノン EOS Kiss Digital X 発表
http://kensei.jugem.cc/?eid=542
EOS Kiss DIGITAL X を販売する Web ページへのリンクあり
（5） デジイチ最新情報！ — 小さく・軽い 10 メガ一眼、キヤノン EOS Kiss ...
http://kensei.jugem.cc/?eid=541
EOS Kiss DIGITAL X のレビュー記事の Web ページへのリンクあり
（6）【ユーリードオンラインショップ】 - デジタルカメラ
http://www.ec-shopping.net/shop/ulead/ default.php/cPath/87_160/sort/2a/page/3
EOS Kiss DIGITAL X は無し。キャッシュにも無し
（7） FC2 News
http://news.fc2.com/?g=13
EOS Kiss DIGITAL X に関する記述無し。キャッシュには EOS Kiss DIGITAL X 発売に関する記述あり
（8） APPLE LINKAGE
http://www.applelinkage.com/
EOS Kiss DIGITAL X に関する記述無し。キャッシュには EOS Kiss DIGITAL X 発売に関する記述あり
（9） トランスウエア、Opera 7 の日本語版パッケージを 3 月に発売 (MYCOM ...
http://journal.mycom.co.jp/news/2003/02/17/15.html
EOS Kiss DIGITAL X に関する記述無し。キャッシュ無し
（10） Mac テクノロジー研究所
http://www.mactechlab.jp/
EOS Kiss DIGITAL X のレビュー記事の Web ページへのリンクあり

図 3 単純なアーキテクチャの場合

（1） 次は 1000 万画素！次は X！ Canon Eos Kiss Digital X 特集
http://www.biccamera.com/bicbic/jsp/w/camera/canon/
EOS Kiss DIGITAL X を販売する Web ページへのリンクあり
（2） 新製品が目白押し !! デジタルカメラがずらりと勢ぞろい ...
http://www.biccamera.com/bicbic/ jsp/w/d_camera/degi_new/index3.jsp
EOS Kiss DIGITAL X を販売する Web ページへのリンクあり
（3） ホームショッピング@Yahoo!オークション店
http://page3.auctions.yahoo.co.jp/jp/auction/c126877263
ヤフーオークションで商売しているショップの EOS Kiss DIGITAL N を販売しているページ
（4） コンパック、オールインワンの低価格ノートを数量限定で発表
http://ascii24.com/news/i/hard/ article/1999/07/01/603126-000.html
EOS Kiss DIGITAL X に関する記述無し。キャッシュには EOS Kiss DIGITAL X のレビュー記事の Web ページへのリンクあり
（5） ソニー、AIBO『ERS-111』を受注販売、今回は数量限定なし、申し込んだ ...
http://ascii24.com/news/i/hard/ article/2000/01/25/606678-000.html
EOS Kiss DIGITAL X に関する記述無し。キャッシュには EOS Kiss DIGITAL X のレビュー記事の Web ページへのリンクあり
（6） EOS Digital への道
http://www.eos-d-slr.net/
EOS Kiss DIGITAL X を販売するオンラインショップへのリンクあり
（7） 豊富な品揃え・天体観測のエキスパート KYOEI ネットショップ
http://www.goto-kyoei.co.jp/netshop/sonota/dengen.html
EOS Kiss DIGITAL X 用のバッテリーを販売するオンラインショップ
（8）【予約】EOS KISS DIGITAL X - デジカメ ゲット！〜楽天デジカメ探しの ...
http://plaza.rakuten.co.jp/dmono/diaryall
EOS Kiss DIGITAL X に関する記述のある日記の Web ページ
（9） マイバッテリープロ 純正バッテリーの 5 倍〜7 倍長持ち
http://www.kyoei-bird.com/jtt/mybattery.htm
EOS Kiss DIGITAL X 用のバッテリーを販売するオンラインショップ
（10） USER'S SIDE USA 日本語版ノート PC/日本語でお買い物
http://www.users-side.co.jp/USA_WEB/ product_pg_usa.php?catagory=digicame&mk=canon&type=1&contury=1
EOS Kiss DIGITAL X を販売するオンラインショップ

図 4 改良されたアーキテクチャの場合

ワード候補の集合と同じであるかのいずれかである。それゆえ、本稿で提案したシステムでは、より良い検索結果が得られるか、あるいは普通のサーチエンジンと同じ検索結果が得られるかのいずれかである。実験の結果から、実際により良い検索結果が得られる場合があることが実証されたので、一つの例の場合しか実験を行っていないが、提案したシステムは有効であると考えられる。

文　　献

[1] Goo, http://search.goo.ne.jp/index.jsp

[2] Google, http://www.google.co.jp/

[3] 山名早人, 村田剛志: 検索エンジンの概要, 情報処理, vol.46, no.9(2005), pp.981-987.