

検索エンジンを利用した信頼性の高い Q&A システム

目本信太郎[†]

[†] 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 検索エンジンは、元々は Web ページを検索するためのものであるが、何かを調べるのに使う人も多い。しかし、元々は Web ページを検索するものであるため、思うような結果を得られないことも多い。また、検索された Web ページに書かれていることはうそを含んでいることもあるので、そのまま鵜呑みにすることもできない。本稿では、検索エンジンと演繹データベースを組み合わせることによって、より信頼性の高い Q&A システムを構築できることを提案する。

キーワード Q&A システム, 検索エンジン, 演繹データベース

A Reliable Q&A System Using Search Engines

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

Abstract Although search engines are originally developed to search Web pages, many people use them to examine something. However, they are originally developed to search Web pages, and people often don't obtain such results as they required. In addition, Web pages of search results often contain something which is not true, then people cannot accept them as they are written. In this paper, we propose that a reliable Q&A system can be made by combine search engines and deductive databases.

Key words Q&Q system, search engine, dedunctive database

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) を知らないとい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、多くのユーザは検索エンジンを使って URL の検索を行なっている。

これは検索エンジン本来の使い方であるが、何かを調べるために検索エンジンを使う人も多い。しかし、検索エンジンは元々は Web ページを検索するものであるため、検索結果に質問の答えについて記述されている Web ページの URL が 1 件か 2 件ぐらいしか含まれておらず、思うような結果が得られな

いことも多い。

そこで、本稿では検索エンジンと演繹データベースを組み合わせることによって、より信頼性の高い Q&A システムが構築できることを示す。すなわち、質問の答えをあらかじめ演繹データベースで導き出し、それをキーワードに含めて検索を行うのである。こうすることによって、質問の答えが記述されている Web ページの URL ばかりが検索結果として返されるようになる。

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、我々が提案する Q&A システムを提示する。第 3 章ではシステムの構築法について述べる。第 4 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. アーキテクチャ

2.1 単純なアーキテクチャ

調べたいことを検索エンジンで調べる場合に、キーワードを選ぶもっとも簡単な方法は、質問から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)。

検索エンジンの Goo [1] では、「富士山の次に高い山は何？」というような日本語で記述された質問でも検索できると謳っているのですが、Q&A システムとして捉えることができるが、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て検索しているだけのようである。これを単純なアーキテクチャと呼ぶことにする。

2.2 改良されたアーキテクチャ

図 2 は自然言語で記述された質問をキーワードに変換する時、あらかじめ質問の答えをキーワードに含めて検索を行う改良されたアーキテクチャの概念図である。検索エンジンのフロントエンドプロセッサは質問の答えを調べるために演繹データベースに対してキーワードの組を渡す。演繹データベースは、「富士山」、「次に」、「高い」、「山」ならば「北岳」などといった演繹ルールを持っている。もちろん、Q&A システムに入力される質問はさまざまなので、演繹データベースはそのすべての質問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。

3. システムの構築法

3.1 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、入力された自然言語で記述された質問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渡す。そして、演繹データベースから返された質問結果を、キーワード候補と一緒にして、キーワードの集合として検索エンジンに渡すようにする。この時、演繹データベースに質問にマッチする演繹ルールがない場合は、キーワード候補をそのままキーワードの集合として検索エンジンに渡す。検索エンジンは渡されたキーワードを使って検索を行う。

3.2 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。2.2 節で例として挙げた、「富士山の次に高い山は何？」という質問の場合、答えは「北岳」なので、「富士山」、「次に」、「高い」、「山」ならば「北岳」、となる。

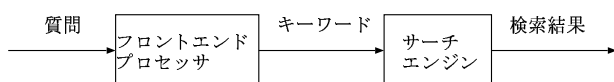


図 1 単純なアーキテクチャ

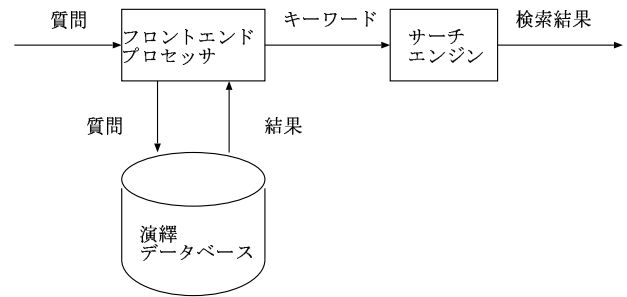


図 2 改良されたアーキテクチャ

記号を使って書き直すと、

”富士山”, ”次に”, ”高い”, ”山” → ”北岳”

と書くことができる。

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。

ところで、演繹ルールを導き出す一般的な方法は自明であるが、一つの方法を述べておくと、例えば「雑学」の本を買ってきて、それに書かれている知識を片っ端から演繹ルールにしてしまう方法が考えられる。また、百科事典を買ってきて、同じようなことができる。演繹データベースの作成者には、特別な知識や技量は特に必要ない。

4. 実行例

2.1 節で例として上げた、「富士山の次に高い山は何？」を用いて実験を行った。

多くの場合、検索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。

Google [2] を使い、「富士山」、「次に」、「高い」、「山」をキーワードとして検索を行った。この場合図 3 のような検索結果が得られた。上位 10 件の検索結果のうち、富士山の次に高い山が北岳であることを述べているのは 6 件であった。

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「富士山」、「次に」、「高い」、「山」、「北岳」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。検索結果のうち、富士山の次に高い山が北岳であることを述べているのは、上位 10 件の検索結果のうち全部で 9 件であった。

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と違って、上位 10 件の検索結果のうち 9 件の Web ページには、富士山の次に高い山が北岳であることが記述されていた。このことから、演繹データベースを用いて、質問の答えをあらかじめキーワードに含めて検索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。

5. 結論

本稿では、検索エンジンを使って何かを調べる場合に、あらかじめ演繹データベースを使って、質問の答えをキーワードに

- (1) 富士山ナビ - Next 富士山 富士山の次に高い山に登る！ 富士山登頂に ...
山頂までの最後の道のりはかなり手強く、恐怖感も高いが、経験が浅くても何とか乗り切れるルート。 ... 季節に応じて初心者から経験者まで楽しむことができるのも八ヶ岳の魅力であり、「富士山の次」にはオススメの山と言えるだろう。 ...
<http://fujisan.yamakei.co.jp/nextfuji04.html>
北岳についての記述あり
- (2) ナンバー 2 の実力ー富士山の次に高い山は？ 編：東京博識倶楽部 ...
ナンバー 2 の実力ー富士山の次に高い山は？」の購入は、新刊・古本の同時購入が可能なオンライン書店の livedoor BOOKS で！1500 円以上のご注文で国内無料配送！さらに、ポイントも貯まります！
<http://books.livedoor.com/item/781569>
北岳についての記述なし
- (3) 富士山 - Wikipedia
富士山は標高は高いが、日本の他の高山に比較すると高山植物などの植生に乏しい。これは富士山が最終氷期が終了した後に山頂から大 親神は次に筑波山の神のもとへ行き、同様に宿を乞うたところ、今度は親神は歓迎された。そのため、筑波山には人々が ...
<http://ja.wikipedia.org/wiki/%25E5%25AF%258C%25E5%25A3%25AB%25E5%25B1%25B1>
北岳についての記述あり
- (4) 富士山の次に高い山はどこですか？ - Yahoo!知恵袋
高い山 日本でいちばん高い山は、富士山らしいですね。では、山頂と登山口の標高差がもっとも大きい山は、どこですか？ また、登るのがもっとも難しいとされる山は、どこですか？(冬山はちょっと除外しましょう) ...
http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q1320962593
北岳についての記述あり
- (5) SATO+LOG:日本で一番高い山、3776m の富士山に登ってみた。
日本で一番高い山、3776m の富士山に登ってみた。 今までそんなに、いや全く気にしなかった富士山ですが、今年は山登り年。赤岳、槍ヶ岳 etc...、いろいろ登っていたら「富士山は登っておくべきだろう」ということになり、富士山へ。 ...
http://blog.livedoor.jp/libra_/archives/51616233.html
北岳についての記述なし
- (6) 山男のつれづれ日記(富士山の次に高い山は?) - elyming 【みんなカラ ...
山男のつれづれ日記(富士山の次に高い山は?)」について - elyming のブログです。Powered by みんなカラ。
<http://minkara.carview.co.jp/userid/132718/blog/298518/>
北岳についての記述あり
- (7) ナンバー 2 の実力 知られざる凄腕たちー富士山の次に高い山は? (ワニ文庫 ...
ナンバー 2 の実力 知られざる凄腕たちー富士山の次に高い山は? (ワニ文庫) に関する情報を一目で調べることができます。販売価格はもちろん、口コミ情報が一目でわかる goo 評判検索です。
<http://buzz.goo.ne.jp/item/cid/9/pcid/7109276/scid/7103/>
北岳についての記述なし
- (8) 富士山の次に高い山・北岳 3192M - くまの花ブログ - 楽天ブログ (Blog)
2008 年 11 月 10 日 ... やっぱ山が好きなので、山の思いで話 ショック になっちゃいます～ うっしっし^^今日は北岳 3192M 富士山 富士山の次に高い山です。 ガンガンやっていた頃は、日帰り登山 3 回 グッド やりました^^一番最初は単独で、登ってみました ...
<http://plaza.rakuten.co.jp/kumamusume/diary/200811100000/>
北岳についての記述あり
- (9) 渋ろぐ：日本一高い山は富士山。二番目は？
2009 年 12 月 9 日 ... 日本一高い山は富士山。二番目は？ 日本一高い山は富士山。 世界一高い山はエベレスト。 それでは日本で二番目に、世界で二番目に高い山は何でしょう？ みなさんは、おわかりでしょうか？(大丈夫です、ほとんどの人が知りません) ...
<http://shibuma.exblog.jp/10542660/>
北岳についての記述あり
- (10) Amazon.co.jp： ナンバー 2 の実力 知られざる凄腕たちー富士山の次に高い ...
Amazon.co.jp： ナンバー 2 の実力 知られざる凄腕たちー富士山の次に高い山は? (ワニ文庫): 東京博識倶楽部: 本. ... ナンバー 2 の実力 知られざる凄腕たちー富士山の次に高い山は? (ワニ文庫) (文庫). 東京博識倶楽部 (編集) ...
<http://www.amazon.co.jp/.../dp/4584304688>
北岳についての記述なし

図 3 簡単なアーキテクチャの場合

(1) 富士山の次に高い山・北岳 3192M - くまの花ブログ - 楽天ブログ (Blog)

2008 年 11 月 10 日 ... ぼっ わからん 曇りかな? やっぱ山が好きなので、山の思いで話 ショック になっちゃいます〜 うっしっし ‘今日は北岳 3192M 富士山 富士山 の次に高い山です。ガンガンやっていた頃は、日帰り登山 3 回 グッド やりました’...

<http://plaza.rakuten.co.jp/kumamusume/diary/200811100000/>

北岳についての記述あり

(2) 富士山ナビ - Next 富士山 富士山の次に高い山に登る! 富士山登頂に ...

なお、日本で 4 番目の高さを誇る間ノ岳 (3189m) も、北岳から 3~4 時間の距離。一挙に日本の最高峰群を制することもできる。 季節に応じて初心者から経験者まで楽しむことができるのも八ヶ岳の魅力であり、「富士山の次」にはオススメの山と言えるだ ...

<http://fujisan.yamakei.co.jp/nextfuji04.html>

北岳についての記述あり

(3) コトノハ - 日本で 2 番目に高い山は?

富士山・北岳・奥穂高岳。 ちぎ. 北岳. anniks. 北岳. 黒ちょこ. 白根山でしょ. ちょっち. 北岳 地元なので小学生のときよく先生が言ってた ... 台湾に 2 番目に高い山は次高山。 鳩山. これはわからん. as365n2. とりあえず〇でも×でもないのは確かだ。 ...

<http://kotonoha.cc/no/57996>

北岳についての記述あり

(4) 山男のつれづれ日記 (富士山の次に高い山は?) - elyming 【みんカラ ...

山男のつれづれ日記 (富士山の次に高い山は?)」について - elyming のブログです。 Powered by みんカラ. ... オフシーズンになるとこの夜叉神峠入口で一般車は通行止めになるので積雪期間の北岳は、アプローチが不便で、ホントに遠い山になってしまい ...

<http://minkara.carview.co.jp/userid/132718/blog/298518/>

北岳についての記述あり

(5) 富士山の次に高い山はこの山ですか? - Yahoo!知恵袋

富士山の次に高い山はこの山ですか? ... 北岳ですね今年行ってきました!! 過呼吸になるのわかるなぁ 頂上で自分も 叫んじゃいました。(が おー) って ちなみに 頂上付近で雷鳥の親子が見れたので ラッキー ...

http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q1320962593

北岳についての記述あり

(6) 北岳北岳紹介

北岳は「百名山」「花の百名山」のどちらにも名前を連ねています。 別名「白根山」とも呼ばれていまして、 標高 3192m の富士山の次に高い山です。 北岳 (右写真には間ノ岳も) 私は 14 歳のときに登ったのですが、 あいにくの悪天候で、霧に包まれていました ...

<http://www.5ksjy5.net/kitadake.html>

北岳についての記述あり

(7) YamaNavi : 山ナビ - 北岳

2 番手の魅力. 富士山の次に高い山となると、その存在は、つまらない事ようになるが北岳はそうではない。その存在は通俗な富士とは別で、気品高く誰もが登ることを好まない、捻くれた面を見ることがある。 ...

<http://www.yamatabinet.com/yamanavi/yama.php>

北岳についての記述あり

(8) 伊豆ネイチャーガイドの独り言:北岳に雪

2009 年 11 月 15 日 ... 北岳に雪 安良里の港から遠望できる南アルプス。ズームで見れば、富士山の次に高い山。「北岳」 その北岳の頂上が白くなっていました。 この写真は 11/3 のものですので、昨日までの天気で甲斐駒から白い峰が並んでいることでしょう。 ...

<http://nature.i-ra.jp/e158026.html>

北岳についての記述あり

(9) 富士の見えるところ、北岳:富士山へ行こう

富士の見えるところ、北岳. 北岳 ここは日本で二番目に高い山、南アルプスの北岳からの富士山です。 広河原から丸一日かけて登りました。 とても疲れましたが、その甲斐あって晴天の富士が見えました。 これは朝早くのもの昼くらいには雲海は消えていまし...

<http://fuji.livedoor.biz/archives/50164398.html>

北岳についての記述あり

(10) マークスの山 北岳ほか (後編) 熟年ヤングハートフォーラム/ウェブリ ...

... 仙丈ヶ岳とこれから向かう間ノ岳。 高い山は県境になることが多いが北岳は単独で山梨県に属する。これから向かう間ノ岳、農鳥岳は山梨と静岡の県境に位置する。 ... さて名前の由来である農鳥岳の雪形は 2 種類あり、夏の初めに白鳥の雪形がでると、イネの苗代を、次に牛の雪形がでるとダイズやアズキを撒いた。 ... 西農鳥岳から見た農鳥岳頂上と富士山、そして 11:35 農鳥岳山頂 (3026m) 着。 バックに鳳凰 3 山 ...

http://30972253.at.webry.info/200709/article_2.html

北岳についての記述なし

図 4 改良されたアーキテクチャの場合

含めて検索を行うような Q&A システムのアーキテクチャについて述べた。実行例では上位 10 件の検索結果のうち 9 件しか質問の答えについて記述されている Web ページを検索できなかったが、理論上は上位 10 件すべての検索結果が質問の答えについて記述されている Web ページであるということも十分ありうる。

読者の中には、質問の答えをキーワードに含めて検索するのだから、そんなこと当たり前ではないのか、と思う人もいるかもしれない。確かに当たり前で、自明のことである。著者はそのことを否定しない。では、当たり前のことをすることにどういう意味があるのかと言うと、質問の答えをキーワードに含めて検索を行うと、答えが確かに質問の答えであることを述べている Web ページが検索されるのである。しかも、上位 10 件のうちのそのほとんどが、である。

演繹データベースを使って質問の答えを提示しただけでは、必ずしも利用者に信用してもらえないかもしれない。しかし、検索エンジンの検索結果は、かなり信用されているそうなので、答えが確かに質問の答えであることを述べている Web ページが 10 件も検索されて出てくれば、多くのユーザはその答えを信用してくれることが期待できる。

また、質問の答えをキーワードに含めて検索を行うと、答えが確かに質問の答えであることを述べている部分が、スニペットに含まれる形で、検索結果が得られることも多い。この場合、検索された Web ページを実際に閲覧することなく、質問の答えを知ることができるので、時間の節約になる。

さらに、検索された Web ページを丹念に閲覧すると、質問の答えの近傍情報が得られる。例えば、「富士山の次に高い山は何？」という質問の場合だと、答えの「北岳」だけではなく、北岳の標高は何メートルであるのかとか、北岳はどこにあるのかとか、北岳の次に高い山はなんと言う山であるのか、といった情報が得られる。

以上のことから、質問の答えをあらかじめキーワードに含めて検索を行う検索エンジンは、信頼性の高い Q&A システムになりうるということが言える。

文 献

- [1] Goo, <http://search.goo.ne.jp/index.jsp>
- [2] Google, <http://www.google.co.jp/>