

検索結果による概念検索システムの有効性の評価

Evaluation of Effectiveness of Concept-based Information Retrieval System

横井 知晴*

Chiharu YOKOI

Abstract

The WWW becomes to contain so much information that a considerable portion of the users has difficulty during the course of information retrieval (IR). In the present paper two cases of IR was compared. One case where a user uses a search engine that utilizes natural language, and the other case where the concept-based IR which analyzes the contents of a query and adds queries derived from the concept of the first query.

The difference in performance of the above two systems is compared. For this purpose, a set of questions that are consisted of a single line scientific documents, either correct or incorrect, are provided. Then those questions are subjected to IR systems. Both systems are designed to align IR results in the order of accuracy of document. Performance was scored as plus marks if the document is sufficient to obtain correct answer otherwise negative marks.

It is shown that concept-based IR is of better performance. Although a question containing the word which has the meaning of a wide sense, performance is worse. A limitation exists that the parent population of the data used in this experiment was dependent on Google, however, it seems that concept-based IR will be helpful to find out answers, at least, to questions; Yes or No.

* 愛知淑徳大学大学院文学研究科図書館情報学専攻

Graduate School of Library and Information Science, Aichi Shukutoku University
JOURNAL OF LIBRARY AND INFORMATION SCIENCE. Vol. 20, p. 51–58 (2006)

1 序論

WWW(World Wide Web)の普及に伴い、ネットワーク上を流通する情報が爆発的に増加した。情報の増加とともに、情報検索(information retrieval: IR)に関して初心者である利用者も増加した。このような利用者の多くは、ネットワーク上に存在する情報から自分に必要な情報を収集し、有効に利用することに困難を感じる情報過多(information glut)という状態にある。ネットワーク上にあるデータをいくら集めても、それはデータでしかなく、それを有効に利用するためには、データから何らかの「知識」などを発見することが必要となる。情報を検索する場合、「ユーザーの検索語に対する知識」、「クエリーを適切な言葉で表現する能力」、「情報探索(検索)経験」などが必要となるものの、それらを有するのは、研究に必要な情報収集を日常的に行っている専門家に片寄っているというのが現状である。専門家は自らの研究に関する知識や、有効な情報源をはじめとする情報探索の手法および知識、それらを利用して収集した情報から適切なものを抽出するための評価基準などを、有している。一方、日常的に情報を収集する事の少ない一般ユーザーがこれらを熟知しているとは、考えにくい。よって、一般ユーザーにとって、情報を検索するという行動は、非常に困難が伴うものであると言えよう。

一般的に情報検索は、検索目的に対する漏れの少なさ(再現性)、ノイズの少なさ(適合性)などによって評価される。これらを十分に得るには、情報検索システムの構造、情報検索システムの使用方法などを熟知し、使用した場合のほうが、高い評価を得られやすい。しかし、一般ユーザーには、このような知識があるとは、言いがたい。

情報検索システムが存在する目的は、大量の情報からユーザーの要求を満たす情報を探索することである。従来のデータベースを例にとると、具体的には、書誌データベースにおいて、

ユーザーが提示するクエリーに適合する書誌データを見つけ出すことである。しかし、情報検索システムを上手に使うためには、「ユーザーの検索語に対する知識などの相違」、「ユーザーとシステム間のクエリー作成に対する解釈の相違」、「検索質問をクエリーに置きかえることが出来ない」などの問題点が存在し、これらが一般ユーザーに困難さをもたらすと考えられる。情報検索システムはキーワードマッチングによってシステムが検索結果を導き出しているため、検索精度は、ユーザーのクエリーに関する知識に左右される。そこで、これらの問題点を解決するには、概念検索(Concept-based information retrieval)が有効であると考えた。

本研究では、WWW上の膨大な文書を検索する際に、「ユーザーの検索語に対する知識などの相違」、「ユーザーとシステム間のクエリー作成に対する解釈の相違」、「検索質問をクエリーに置きかえることが出来ない」という問題点を改善するためにユーザーが自然言語を用いて検索エンジンを用いた場合と、クエリーの内容を解析し、概念を加える概念検索を用いた場合を取り上げ、両者のパフォーマンスの違いを比較し、どちらが有効であるかを明らかにすることを目的とする。

現在の情報検索システムは検索質問中の語句と文書中の語句との文字列としての一致を基本としている。したがって、漏れのない網羅的な検索を実行しようとするれば、同義語や言い換えを列挙し、OR(論理和演算子)で連結する必要がある。しかし、これは実際のユーザーには難しく、負担となるので、語句ではなく、その上のレベルの「概念(クエリーやキーワードの持つ意味的な情報)」に基づく検索に関する研究が進んでおり、このための工夫として、シソーラスなどの統制語彙が以前より開発されてきた。このような道具をユーザーが明示的に使うのではなく、システムによる自動的な概念検索を実行可能とすることが近年の研究の焦点である。

概念検索(Concept-based information retrieval)

とは、データベースに検索条件を語句で入力するのではなく、文章あるいは複数の語を与え、それらとデータベース中のデータに入力された語の出現頻度で構成される概念の類似度を算出し、文書を探し出す検索手法である。この手法では、回答について類似度を算出し、類似度の高いものから順に回答が表示される。近年、ビジネス分野、特にキーワード設定が困難と言われるビジネスモデル特許の検索において注目されている技術の一つである。

概念検索は、類似検索 (Query by Example)、類似文書検索という表現もされているが、それらも、概念検索の一種である。この技術を使うと、同一の単語 (共通単語) が含まれていなくても、概念的に近似の文書を探し出すことができ、的確な文書を入手する可能性が飛躍的に高まる。

概念検索は、従来の検索方法とは以下の点で大きく異なる。

- (1) 検索式を作らずに文章で検索することができる。
- (2) 文章を入力せずに類似文章を検索することができる。
- (3) あらかじめ人手で辞書やソーラスを用意しておく必要がない。

概念検索の長所としては、関連データベースの保守の必要がない、人間では思いつかないような関連付けを作ることが可能なことがある。一方、短所としては、検索対象の文書群に検索結果が大きく依存する、目的に合わない文書を多く検索する、つまりノイズを増やす可能性があることである¹⁸⁾。

概念検索に関する研究は、大きく7つに分けることが出来る^{23),24)}。

- (1) 概念検索の動作原理
- (2) 検索できるもの、できないものの内容の相違
- (3) ランキング方法および、ランキングの順位
- (4) 再現率および、精度
- (5) 質問文の妥当性
- (6) 絞込み検索や検索式の併用の利用法

(7) 特定の知財関連業務に用いる際の的確性

概念検索では、データ集合にある文書の内容を対象にして入力された文の内容に関連する文書に関連性の高い順に抽出する。キーワード検索では、入力されたキーワードと一致するレコードを抽出する。また、全文検索では、文書の内容を対象に入力されたキーワードが含まれる文書を抽出する。クエリーの例としては、概念検索では、「A社から発売されているテキスト検索の製品について知りたい」であるのに対して、キーワード検索では、会社コード、製品No.を入力する必要がある、全文検索では、("A" or "B") and ("C" or "D") といった論理演算子を用いてキーワードを表現する必要がある。

2 実験

2.1 実験方法

本実験は、以下の2点の仮説に基づいて行った。

- (1) Googleでは、語句を用いて検索を行うが、検索語がユーザーの知識や検索によって変わる。一方、概念検索システムでは、自然文を用いるため、検索結果に差が出ると予想される。
- (2) 概念検索システムを用いて検索を行うことにより、ユーザーのクエリーに対する概念に対して、科学的思考をもとにした適切な補助を与え検索することが、可能になると予想される。

質問項目を形態素解析システムである「茶筌 version2.1 for Windows」を用いて形態素を抽出し、すべての形態素を検索語として用いた。ただし、Googleには、表記のゆれが存在するため、表記のゆれに配慮する必要がある。情報検索を行う際、名詞が最も使われやすい。よって、表記のゆれに配慮し、形態素解析によって得られた名詞句は、考えられるあらゆる表記方法でGoogleを用いて検索を行った。ただし助詞、助動詞は、Googleでの検索では、用いら

ない事とした。これらを用いることにより、検索結果にノイズが増え、その結果不要文書が増えるからである。このようなストラテジーのもとで、Googleで検索された、最大Top30までで構成されるデータのデータ集合を作成した。そのデータ集合をGoogleでの検索結果とし、同じデータ集合を概念検索システムにおいて自然言語で検索した検索結果を概念検索システムでの検索結果とした。

2. 2 質問項目

本実験で用いる質問項目(情報要求)は、科学的思考を持つという仮説を採択するため、『月刊化学』で毎月特集されている「ぼくらの科学リテラシー」の質問項目を用いてGoogleで検索したものを実験データとして用いた。ここで用いた質問項目は、2004年の1月から5月までに月刊化学に出題され、その正解も掲載されたものを用いた。

本研究で用いた質問項目を以下に例示する。

- (1) 砂糖をかけてもナメクジは縮む。
- (2) サンゴ礁のサンゴは動物である。
- (3) サボテンのトゲは葉が進化したものなので、気孔が存在する。

科学リテラシーで取り上げられている質問項目は、全て正解、不正解で答えられる質問である。このような条件下では、答えが不正解の質問項目には、ユーザーの検索語に対する知識などの相違、ユーザーとシステム間のクエリー作成に対する解釈の相違という問題点の両者をあらかじめ含んでいることになる。正解、不正解の質問項目を混在させることにより、検索エンジンを用いた場合と、クエリーの内容を解析し、概念を加える概念検索の場合の両者のどちらが有効であるか否かが明らかになるであろう。

2. 3 本研究の情報検索システム

本研究で用いる情報検索システムは、Googleと概念検索システム(Concept Search)を用い

る。本研究で用いた概念検索システムは、自然言語で入力し、類似度の判定には、統計処理を行い、検索結果を出力させている。自然言語の解析には、形態素解析を用いている。また、統計処理には、ベクトル空間モデルを用いているシステムである。

2. 4 評価尺度

また、本研究で用いた質問項目は、月刊化学で毎月特集されている「ぼくらの科学リテラシー」の質問項目であり、正解は、質問掲載号の発行された翌月に、月刊化学のWebサイトおよび雑誌で科学的根拠の解説を含めて公表されており、本実験での評価基準として用いるために必要な、信頼性と正確性を備えていると考えられる。公表された回答結果に基づいて、評価を行う。以下に評価尺度を示す。

- A: 質問項目の記載があり、なおかつ適切なもの(適合文書)
- B: 質問項目が記載してあるものの、記載事項に相違があるもの
- C: 質問項目の記載が無く、不適切なもの

2. 5 評価方法

得られた検索結果を、一つずつブラウジングを行い、評価尺度である質問項目の記載の有無を点数化し基本統計量を求めて評価を行う。

点数化の方法は、最上位にランク付けされたものが検索結果数の値(検索を行い、得られた検索結果が30件ならば、最上位は、30点)で、以下ランクが一つ下がれば1点ずつ減っていき、最下位にランク付けされたものは、1点として評価する。また、記載されていない、つまり適合文書ではない文書には、負数を与える。ただし、記載がされているものの、記載内容の異なる文書に関しては、点数を半分にして与える。

また、Top10の検索結果を用いて出現順位の差(本論文では、オッズと呼ぶこととする)を測り、基本統計量を求めて評価を行う。出現順

位の差を求める算出方法は、まず、概念検索システムのTop30とGoogleのTop30の差を算出する。そして、その差のTop10を対象に総和を算出し、その総和を質問項目の出現順位の差とする。

3. 実験結果

表1は、実験結果の一例である。順位とはGoogleあるいは概念検索により表示された順序を検索結果の順位としたものである。DNとは、Googleによる検索結果の順位を基に、その順位をそのままドキュメントナンバーとして付与した番号である。2列目と6列目のA, B, Cとは、評価結果であり、Aは評価尺度の記載があったもの、Bは記載があったものの内容が異なっていたもの、Cは、評価尺度の記入の無かったものである。

表1 実験結果の一例

(1) サンゴ礁のサンゴは動物である。

順位	評価	Google	順位	概念検索	評価	DN	順位差	オッズ
1	A	30	1	30	A	5	4	
2	C	-29	2	29	A	27	25	
3	A	28	3	28	A	22	19	
4	A	27	4	27	A	18	14	
5	A	26	5	26	A	15	10	
6	C	-25	6	25	A	20	14	
7	A	24	7	24	A	12	5	
8	A	23	8	23	A	1	-7	
9	A	22	9	22	A	7	-2	
10	A	21	10	21	A	13	3	85
11	A	20	11	20	A	21	10	
12	A	19	12	19	A	8	-4	
13	A	18	13	18	A	14	1	
14	A	17	14	17	A	26	12	
15	A	16	15	16	A	17	2	
16	C	-15	16	15	A	28	12	
17	A	14	17	-14	C	25	8	
18	A	13	18	-13	C	16	-2	
19	A	12	19	12	A	23	4	
20	A	11	20	11	A	19	-1	
21	A	10	21	-10	C	29	8	
22	A	9	22	9	A	30	8	
23	A	8	23	8	A	11	-12	
24	C	-7	24	-7	C	2	-22	
25	C	-6	25	-6	C	6	-19	
26	A	5	26	-5	C	24	-2	
27	A	4	27	4	A	9	-18	
28	A	3	28	3	A	4	-24	
29	C	-2	29	2	A	3	-26	
30	A	1	30	1	A	10	-20	
総合計								0

表2 質問11項目全データの基本統計量

	Google	概念検索
平均	-12.85	-12.50
標準誤差	0.46	0.48
中央値	-14.00	-14.00
最頻値	-27.00	-30.00
標準偏差	12.75	13.14
分散	162.46	172.66
尖度	1.81	1.59
歪度	1.22	1.20
範囲	60.00	60.00
最小	-30.00	-30.00
最大	30.00	30.00
合計	-10052.00	-9414.00
標本数	782.00	753.00

また、この実験の質問11項目に関する全データの基本統計量を表2に例示する。これらの11項目のTop10, Top20, Top30, 質問項目に広義の意味を含む語を含むもの、質問項目に狭義の意味を含む語を含むものに分け、それぞれ基本統計量を求め、パフォーマンスの差があったか否かを求めた。以下に実験結果の詳細を示す。

(1) 11項目のTop10, Top20, Top30

表3のようにTop10では、Googleの実験結果は平均が-7.37, 標準偏差が24.79であり、概念検索システムの結果は、平均が-1.30, 標準偏差が25.84であった。なおTop20では、Googleの実験結果は平均が-14.1, 標準偏差が16.0であり、概念検索システムの結果は、平均が-13.0, 標準偏差が16.9であった。Top30では、Googleの実験結果は平均が-10.9, 標準偏差が14.0であり、概念検索システムの結果は、平均が-10.3, 標準偏差が14.5であったので、パフォーマンスの差が見られた。

表3 Top10 の基本統計量

	Google	概念検索
平均	-7.37	-1.30
標準誤差	3.20	3.34
中央値	-22.50	-21.00
最頻値	-27.00	28.00
標準偏差	24.79	25.84
分散	614.47	667.94
尖度	-1.59	-2.00
歪度	0.64	0.13
範囲	60.00	60.00
最小	-30.00	-30.00
最大	30.00	30.00
合計	-442.00	-78.00
標本数	60.00	60.00

(2) 質問項目に広義の意味を持つ語を含むもの
 Top10では、Googleの実験結果は平均が-15.9、標準偏差が20.3であり、概念検索システムの結果は、平均が-10.6、標準偏差が23.6であった。Top20では、Googleの実験結果は平均が-13.9、標準偏差が16.2であり、概念検索システムの結果は、平均が-11.1、標準偏差が18.2であった。Top30では、Googleの実験結果は平均が-10.0、標準偏差が14.7であり、概念検索システムの結果は、平均が-8.5、標準偏差が15.6であったので、パフォーマンスの差が見られた。

(3) 質問項目に狭義の意味を含む語を含むもの
 Top10では、Googleの実験結果は平均が-15.7、標準偏差が20.5であり、概念検索システムの結果は、平均が-19.3、標準偏差が17.0であった。Top20では、Googleの実験結果は平均が-15.6、標準偏差が14.6であり、概念検索システムの結果は、平均が-15.9、標準偏差が14.2であった。Top30では、Googleの実験結果は平均が-12.0、標準偏差が13.1であり、概念検索システムの結果は、平均が-12.4、標準偏差が12.7であったので、パフォーマンスの差が見られた。

また、出現順位の差は、広義の意味を持つ語を含むものは、平均が77.4、標準偏差が31.97であり、狭義の意味を持つ語を含むものは、平

均が66.8、標準偏差が13.17であった(表4)。この結果、広義の意味を持つ語を含むものには、ばらつきが現れ、狭義の意味を持つ語を含むものは、有効な文書が上位に出現する事が分かる。

表4 広義の語と狭義の語を含むものの基本統計量

	広義	狭義
平均	77.40	66.83
標準誤差	14.30	5.38
中央値	94.00	69.50
標準偏差	31.97	13.17
分散	1022.30	173.37
尖度	-3.03	-0.41
歪度	-0.44	-0.21
範囲	67.00	37.00
最小	43.00	48.00
最大	110.00	85.00
合計	387.00	401.00
標本数	5.00	6.00

概念検索システムの場合、狭義の意味を持つ語を含むものは、上位に適合文書が出現した結果から、上位十件を対象に順位の変化の差を測るため、基本統計量をさらに求めた。また、順位の変動を比較するため、オッズを算出した。オッズの算出方法は、Googleでの順位と概念検索システムでの順位の差を算出するものとする。そして、そのオッズの基本統計量を算出し、変動を評価することとする(表5)。

表5 オッズの基本統計量

平均	64.82
標準誤差	6.48
中央値	68.50
標準偏差	34.30
分散	1176.37
尖度	-0.29
歪度	-0.20
範囲	134.00
最小	0.00
最大	134.00
合計	1815.00
標本数	28.00

4. 考察

本実験の標準偏差から、11項目のTop10, Top20, Top30, 質問項目に広義の意味を含む語を含むもの、質問項目に狭義の意味を含む語を含むもの全てにおいてパフォーマンスの差が現れた。また、上位に適合文書が集合し、下位に不要文書が集合していることが分かる。また、平均値から、Googleより概念検索システムのほうが適合文書上位に位置し、不要文書が下位に位置するという結果が得られた。

本実験で用いたデータの母集団は、Googleに依存しているという制限が存在する。しかし、本実験において、質問項目に抽象語つまり曖昧な語が存在する場合、概念検索システムを用いても、適合文書を探し出すことは、不可能であった。一方、質問項目に抽象語が無い場合は、ほとんどの質問項目において概念を拡張することによって、良い結果が得られた。

Googleでは、語句を用いて検索を行い、概念検索システムでは、自然文を用いて検索を行った結果、検索結果に大きな差が出た。Googleの場合、質問項目が間違っただけであれば、Googleで探し出すことは不可能であり、ノイズを増やす一因であり、全く異なる文書が上位に来る。一方、概念検索の場合は、同じ分野であり、関連はするものの、質問項目が記載されていないものが上位に来る結果が得られた。これは、概念検索システムを用いて検索を行うことにより、ユーザーのクエリーに対する概念に対して、科学的思考をもとにした適切な補助を与え検索することが、可能になったからであると考えられる。

また、広義の意味を含む語を用いて検索を行うと、Googleでの検索結果は、膨大な数の文書が玉石混淆でヒットする。広義の意味を含む語とは、本実験の実験項目では、夏や冬、乾燥などである。これらの語では、様々な語と関連しユーザーの意図する事を理解することが困難であるからであると考えられる。

しかし、狭義の意味を含む語を用いて検索を

行うと絞り込まれた適合文書がヒットするという結果が得られた。狭義の意味を含む語とは、本実験の実験項目では、ナメクジやサンゴ、サボテンなどである。ナメクジとは軟体動物、サンゴとは腔腸動物、サボテンとは植物であるということが一般的に知られているため、ユーザーの意図する事を理解することが容易であるからであると考えられる。

5. 結論

概念検索では、自然文で検索を行うことが可能である。また、検索に必要な適切な補助を与えることが可能であるため、調べ学習など学習ツールの一つとして用いることが期待できる検索方法であるといえる。

また、本研究では、Googleでの検索結果をもとに次の段階の実験を行った。概念検索を適用しても、あまり良い結果の得られなかった例では、インターネット上に膨大な数の類似したデータが存在し、最初のGoogleによる検索で10万件以上のヒット数があることもしばしばであった。この膨大なデータのうちの一部をもとに概念検索を行うのではなく、例えば、広告、知識といった分野分けをもとに、インターネット上の情報を選別し整理する必要がある。

謝辞

研究を進めるにあたり、懇切丁寧なご指導をいただいた愛知淑徳大学図書館情報学科の林博司教授、西荒井学教授、三和義秀教授に深く感謝の意を表す。

【参考文献】

- (1) Noh, Young-Hee. A study on the estimation of performance of the concept-based information retrieval model for searching the Web. Journal of Information Science. Vol.28, no.5, 2002, p407-415.

- (2) Efthimiadis, Efthimis N. Query Expansion. Annual Review of Information Systems and Technology (ARIST). v31,1996, p 121-187.
- (3) 岸田和明. 情報検索の理論と技術. 東京,勁草書房,1998, 314p, (図書館・情報学シリーズ3) .
- (4) Parsaye, Kamran. et al.(近谷英明訳)知的データベース.東京,オーム社,1992. 544p.
- (5) 北研二ほか. 情報検索アルゴリズム. 東京, 共立出版, 2003. 212p .
- (6) 首藤公昭. 日本語の形態素解析について. 計算言語学, vol.21, no.3, 1980, p.1-6.
- (7) 野村直之. Concept Baseの言語処理と新しいソリューション. 情報処理学会研究報告, vol.129, no.1, 1999, p.1-8
- (8) 相馬隆宏.概念検索が可能な検索エンジン. 日経オープンシステム, vol.67. 1998, p.108-111.
- (9) 松本裕治. 形態素解析システム「茶釜」. 情報管理, vol.41, no.11, 2000, p.1208-1214.
- (10) 高橋昭公. 特許情報における主題情報の探索: 概念検索とその限界. 情報の科学と技術. vol.54, no.7, 2004, p.355-362.
- (11) Taghva,Kazem. et al. The role of manually-assigned keywords in query expansion. Information Processing and Management. Vol.40, 2004, p.441-458.
- (12) 岸田和明. 文書検索におけるクエリーの拡張方法:大域的解析と局所的分析の実証比較. 情報処理学会研究報告, Vol.2001, No.067, 2001, p.55-62.
- (13) Khan,M, Shamim; Khor,Sebastian. Enhanced web document retrieval using automatic query expansion. Journal of the American Society for Information Science and Technology, Vol.55, No.1, 2004, p.24-40.
- (14) Parsaye, Kamran et al. Intelligent Databases. New York, John Wiley & Sons, 1989, 479p.
- (15) G G Chowdhury. Introduction to Modern Information Retrieval. 2nd. London, Faced Publishing,2004, 474p.
- (16) Alemayehu. Nega. Analysis of Performance Variation Using Query Expansion. Journal of the American Society for Information Science and Technology. vol54, no.5, 2003, p.379-391.
- (17) 野村総合研究所.概念検索 [online] .2006 [cited 2006-1-25]. Available from World Wide Web: <<https://www.patent.ne.jp/01gaiyo/s-point/06.html>> .
- (18) Canon. Canon Technology: 文書概念検索技術 [online]. 2006[cited 2006-1-25]. Available from World Wide Web: <<http://www.canon.co.jp/technology/software/09.html>> .
- (19) G-Search.概念検索の利用方法 [online] . 2006 [cited 2006-1-25]. Available from World Wide Web: <http://db.g-search.or.jp/wdbs/help/WATM/gainen_kensaku_koho.html> .
- (20) Just System. Concept Baseテクノロジー: Concept Baseのできるこゝ. [online] . 2006 [cited 2006-1-25]. Available from World Wide Web: <http://www.justsystem.co.jp/km/whats/search_q_103.html>.
- (21) 化学.東京,化学同人,2004-.月刊.
- (22) 福島俊一.WWW情報検索技術と評価の問題. 情報管理. vol.41, no.8, 2000, p.913-916.
- (23) 六車正道. 概念検索システムの現状と使いこなしの検討(上): 知財力強化に貢献する概念検索. 発明, vol.102, no.4, 2005, p.65-71.
- (24) 六車正道. 概念検索システムの現状と使いこなしの検討(下): 知財力強化に貢献する概念検索. 発明. Vol.102, no.5, 2005, p.45-53.
- (25) 六車正道. 概念検索による特許情報の活用方法(1). 発明. Vol.100, no.4, 2003, p.52-56.
- (26) 六車正道. 概念検索による特許情報の活用方法(2). 発明. Vol.102, no.5, 2003, p.45-53.
- (27) 六車正道. 概念検索による特許情報の活用方法(3). 発明. Vol.102, no.6, 2003, p.50-59.