共起単語情報から検索質問ベクトルを補完する用語検索エンジンの実装

田内 学　Nigel Ward
東京大学 工学部 機械情報工学科
{manabu,nigel}@sanpo.t.u-tokyo.ac.jp

1　はじめに

近年、インターネット上に構築されているWWWは、あらゆる種類の膨大な情報を持っており、最新の情報を提供してくれる。我々は居ながらにしてこれらの情報を自由にアクセスできるようになっており、情報獲得にWWWを利用する機会が増えてきている。WWW上での情報検索に、検索エンジンが良く利用されるが、その際、検索結果を絞り込むために、適切な検索質問を設定しなければならず、多くの場合、何回かの試行錯誤を必要とする。

ある用語について調べたいときに、辞書や事典の代わりに検索エンジンを利用して、その用語に関する説明の記述があるページを探すことが多いが、こういった理由から、その用語以外にも適切な索引語を設定しないと目的の情報を得ることができないといった問題があり、目的の情報を獲得するのが大変困難である。そこで、本研究では、WWWの辞書的利用を目的とした用語検索において、共起単語情報から検索質問を補完して検索を行う用語検索エンジンを実装した。

用語説明として、用語の定義を検索する目的の用語検索は、桜井らのパターンマッチングによる用語検索[1]があるが、本研究では、用語の定義に限定せず、その用語をテーマとしている記述の検索を目的にしている。

2　検索質問の拡張

用語検索のように、索引語が一つしかないとき、検索質問が不足するため、検索質問を補う必要がある。索引語を含んでいるWebページ集合において、出現頻度の高い語は索引語に関連があると考えられる。

2.1　関連研究

このような単語の共起情報を利用して検索質問を拡張する方法にBuckley'sのpseudo-feedback[2]がある。pseudo-feedbackは、2段階検索によって、短い検索質問を拡張する方法で、1回目の検索は元の短い検索質問で行い、その結果から、以下の式で検索質問を拡張して2回目の検索を行うものであり、TREC4で有効性が示されている。

$$\overrightarrow{Q_{expanded}} = \overrightarrow{Q_{original}} + average\{\vec{d_i} | i \in (\text{検索結果})\}$$

2.2　拡張検索ベクトルの作成

本システムでは、検索用語を含むページ集合の平均ページベクトルが、その検索用語の性質を表していると考えた。WWWではページによって文書ベクトルの大きさがかなり異なるので下式でのd_iはベクトル長を1に正規化した文書ベクトルである。

$$(DTf_1, DTf_2, DTf_3, ..., DTf_n, ...) = average\{\vec{d_i} | i \in (\text{検索結果})\}$$

DTf_n :単語nの検索結果集合における頻度

検索結果集合のベクトルの元である各単語と索引語との関連性が高いほど、一般的なページに比べて、検索結果集合に多く生起していると考えられる。そこで、これを利用して、検索用語との関連度を次式のように定義する。

$$R_n = \max(\log(D_n/G_n), 0)$$

R_n :単語nの検索用語に対する関連度

D_n :検索結果集合における単語nの生起確率 $(DTf_n/\sum DTf_i)$

G_n :WWW一般における単語nの生起確率

大きな重みを持つ一つもしくは少数の単語に、左右されないようにするため、対数によって重み付けを行うことにした。1ページのみに生起して、他のページに生起しない単語はそのページに偏って生起した単語と考えれるため、そのような単語の重みを全て0とした。

この関連度R_nを各単語の重みとして、拡張検索ベクトル$\overrightarrow{ExQ}$を$(R_1, R_2, \cdots, R_n, \cdots)$とする。

3　Webページの評価

Webページを評価するファクターには、ページの内容、リンク構造による引用関係、ページの更新日時、

アクセス数などいろいろあり、商用検索エンジンはこれらのファクターの一種類または複数種類を利用している [5]。本システムのように、用語の意味情報を持つページを検索する目的においては、これらのファクターのうち、特にページの内容が重要であると考えられるので、それ以外のファクターについては考慮しなかった。

内容の評価は、拡張検索ベクトルを用いて次の2つの検索モデルによって行った。

3.1 ベクトル空間法

一般的な検索モデルであるベクトル空間法を利用して、正規化された各ページベクトルと拡張検索ベクトルの内積から適合度を求める。これは、語の生起頻度に着目して求めた適合度を表しており、ページの方向性が検索質問にどれだけ近いかを表している。

$$VSS_i = \overrightarrow{d_i} \cdot \overrightarrow{ExQ}$$

VSS_i :ページ i のベクトル空間法による適合度

3.2 関連文集合度

関連文集合度によって、検索対象ページが検索用語についての説明的な記述があるかどうかを調べる。

Webページは多種多様で、文体も様々である。表とかリストのようなものから、新聞記事、論文のような説明的な文章で書かれているものまである。辞書的に検索エンジンを利用する際に、多くの場合求められる文体は、説明的な文章であり、関連度の高い語が多く共起する関連文である。

そこで関連文の尺度として関連語共起度を考えた。

関連度共起度は下式で求める。ただし、一つの文に同じ単語が複数生起するときは、1回生起したものと同じと考える。なお、文は、清田らのWWWからの文抽出のルール [3] のHTMLタグによって区切り、その後、日本語形態素解析システム『茶筌』[4] によって文法的に文を区切る

$$CR_k = \frac{\sum_i \sum_{j,i \neq j} R_i \cdot R_j}{{N_k}^2}$$

CR_k :文 k における関連語共起度
R_i :文 k に生起する単語 i の索引語関連度
N_k :文 k を構成する単語の数

また、高い関連語共起度を持つ文が連続しているほど、より多くの情報をもつと考えられるので、関連文の集合度を計る尺度として関連文集合度を考えた。関連文集合度はこの関連語共起度を利用して、以下の式で定義する。

$$LD_i = \max\{\sum_k CR_k \times \max(10 - |x-k|, 0)| \quad 0 \leq x \leq n_i\}$$

LD_i :ページ i の関連文集合度
n :ページ内の文の数

図1は先頭から k 番目の文の関連語共起度を表しているグラフの例で、これに図2のようなフィルターをかけたときの最大値が関連度集合度である。この例では、ページの真中付近に関連文が密集しており、この付近にフィルターの中心があるときが最大になり、その値が関連文集合度となる。フィルターの中心から前後10文について、関連文集合の対象としているが、これは実験的に選んだ。

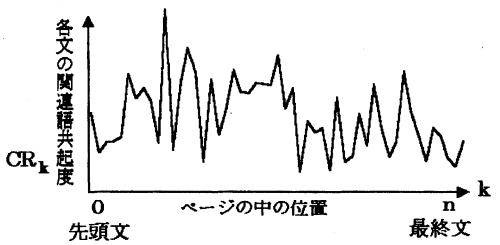

図 1: あるページにおける各文の関連共起度の例

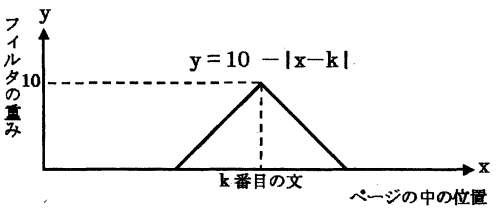

図 2: 関連文集合度を測るためのフィルター

4 用語検索エンジンPerrie

検索エンジンPerrieは、図3の流れで処理を行なっている。

検索結果の取得はInfoseekを利用して上位100件を取得する。次に形態素解析システム『茶筌』により、取得したページから名詞を抽出し、そのデータから、§2.2の拡張検索ベクトルを作成をする。そのベクトルを基に§3.1のベクトル空間法による適合度、§3.2の関

連文集合度によって各ページの評価を行い、ページの総合評価を行う。ユーザーの求められるページは、両尺度が高いページであると考えられるので、両尺度の積から評価得点を算出する。

$$P_i = LD_i \cdot VSS_i{}^{\alpha}$$

P_i :ページ i の評価得点
α :定数

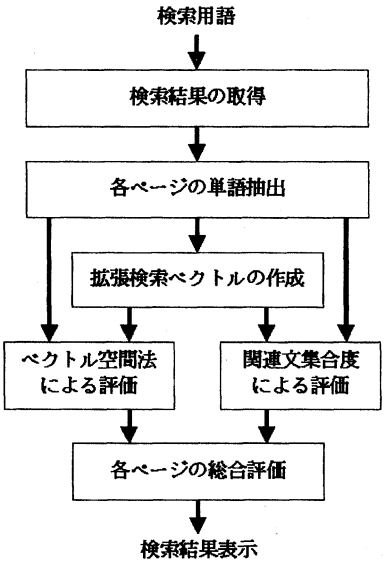

図 3: 検索エンジン Perrie の全体フロー

5 評価実験

5.1 ランク付け検索システムの評価方法

従来、ランクキング検索システムの評価には、適合率、再現率が利用されている。この方法は文書が検索目的に適合するか、適合しないかに二分できるという前提に基づいている。

しかし、Web ページは多種多様であり、一概に、適合、非適合に分類できないし、同じ適合しているページでも、そのページの持つ情報に優劣がある。そのため、多段階評価が必要である。さらに、WWW を扱う検索エンジンにおいて求められるのは、多くの情報を持つ適合ページに早くたどり着くことであり、特許検索のように全ての適合ページが求められるわけでなく、どれか一つでよいことが多い。そのため、適合率が求められ、再現率はあまり求められない。これらの理由から新しい評価方法が必要である。そこで、各ページを 7 段階評価し、その得点によってランク付け検索システムの評価を行う、新しい評価方法を提案する。

高得点のページをなるべく上位に表示するランク付けが良いランク付けであると考えられるため、評価得点は以下の式で算出する。

$$(\text{ランク付けの評価得点}) = \frac{(i\text{番目の出力結果の点数})}{\log(i+1)}$$

この得点から、ランキング精度を以下の式で算出する。

$$\frac{(\text{ランク付けの評価得点}) - (\text{評価得点の期待値})}{(\text{評価得点の理想値}) - (\text{評価得点の期待値})}$$

ここでの、評価得点の理想値とは、各ページをユーザーの評価順に並べた場合の評価得点を指し、評価得点の期待値とは、無作為にランク付けを行った際の評価得点の期待値を指す。この式によって、無作為なランク付けのランキング精度は 0 に、理想的なランク付けを行う検索エンジンのランキング精度は 1 に正規化される。

5.2 実験方法

被験者に以下の質問をした。「これから、ある語の意味やその語に関する意味情報を得る目的で辞書的にインターネットを利用して用語検索を行います。何か一つ調べたい語を挙げて下さい。」挙げられた語を索引語にして検索エンジン Infoseek で検索を行い、その検索結果上位 40 件について、7 段階（1. 全く関係ないページ/全く役に立たないページ～7. 非常に満足できるページ）の評価をしてもらった。評価は、そのページ自体の内容の評価であり、リンクサイトとしての評価は除いてもらった。

5.3 結果

実験では、以下の 19 個の検索用語が挙げられた。

IT 革命, 出師表, ADHD , 関羽, 屈原, 直木三十五, ユーゴスラビア, マイライン, 行為障害, オギノ式, 筋ジストロフィー, ニューディール政策, ラマーズ法, 纏足, 金庫株, ワルサー P38 , スウィングバイ, ハプスブルグ家, 宮部みゆき

この 19 個の検索用語それぞれについて、被験者に評価してもらった 40 件の Web ページを以下の A～D の 4 つの手法で、ランク付けを行った。そのランク付けを 4.1 で定義した評価方法で評価し、ランキング精度の比較を行った。評価得点、ランキング精度の計算はランク付け結果のうち上位 20 件を対象にした。定数 α は 1.0 とした。

A. 従来の方法 pseudo-feedback(§2.1) によって検索ベクトルを拡張し、ベクトル空間法により求めた適合度によってランク付けを行う手法

B. Perrie の算出した拡張検索ベクトル (§2.2) を使って求めたベクトル空間法 (§3.1) の適合度によるランク付け手法

C. Perrie の算出した拡張検索ベクトルを使って求めた関連文集合度 (§3.2) によるランク付け手法

D. 検索エンジン Perrie (B & C の総合評価)

各検索用語についての Infoseek 及び A,B,C,D の手法によるランキング精度は図 4 のようになった。手法同士の有意差検定を行ったところ、BC 間以外は全て有意差があった。

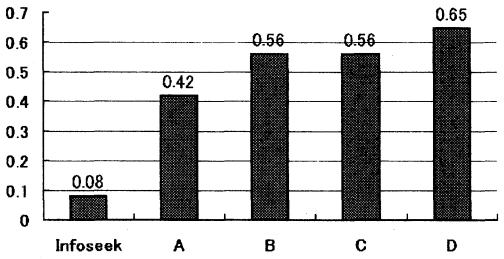

図 4: 各手法のランキング精度の平均値

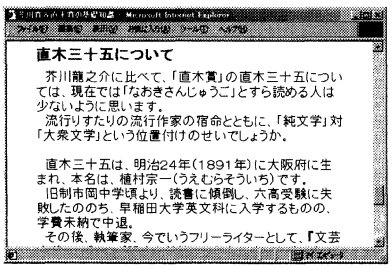

図 5: 検索用語 “直木三十五” ランキング 1 位のページ

6 考察

図 4 からも、見てわかるように、WWW 上の平均的な語の頻度を利用して拡張検索ベクトル (B) を作ることで、pseudo-feedback の従来の手法 (A) よりも、ランキング精度において平均で 0.14 ポイントの有意な上昇が見られ、索引語 19 語中 15 語においてポイントの上昇が見られた。これにより、本研究の拡張検索ベクトルの有効性が示すことができた。

次に、拡張検索ベクトルを使用した B、C、D の手法を比較してみる。ベクトル空間法による手法 B と、関連文集合度による手法 C では有意差は見られないが、両者を使った手法 D (Perrie) では、ベクトル空間法だけを使った手法 B に比べて、平均 0.09 ポイント、ランキング精度が上昇して、有意差が見られた。これは、2 つの検索モデルを用いることで、2 種類の “悪いページ”(被験者の評価が低かったページ) をランクの下にまわすことができたことによると考えられる。

ランキング精度が低くなった索引語の原因としては、次の 3 つが考えられた。

- 検索結果集合が複数のクラスタから構成されている
- 見せかけの相関から関連度が正しく求まらなかった
- 被験者のページ評価に画像等の視覚情報も含まれた

本論文で実装した検索エンジン Perrie は Infoseek と比べて、19 語中 18 語でランキング精度が上がり、平均では 0.57 上がっていた。また、図 5 の例のように、19 語中 14 語で Perrie の検索結果上位 3 件に、被験者が最高得点をつけたページが含まれていた。

7 結論

用語検索において、本研究で提案した拡張検索ベクトル、及び関連文集合度によって、より有効な Web ページ評価ができることが示せた。また、実験結果から、WWW の辞書的利用に有効な検索エンジンを実装することができたといえる。

参考文献

[1] 桜井裕, 佐藤理史 “ワールドワイドウェブを利用した用語検索の実現”, 情報処理学会研究報告, 2000-NL-137-4,2000.

[2] Chris Buckley, Amit Singhal, Mandar Mitra, and Gerard Salton: “New Retrieval Approaches using SMART:TREC4”, *The Fourth Text REtrieval Conference(TREC-4)*, In D.K.Harman,1996.

[3] 清田陽司, 黒橋禎夫: “WWW テキストの自動要約と KWIC インデックスの作成”, 情報処理学会研究報告, 2000-NL-137-5,2000.

[4] 松本裕治, 北内啓, 平野善隆, 山下達雄, 浅原正幸: 日本語形態素解析システム『茶筌』version2.0 使用説明書 第二版, 奈良先端科学技術大学院大学,1999.

[5] 福島俊一: “WWW 情報検索技術と評価の課題”, 情報処理学会誌,Vol.41,No.8,pp913–916,8 2000.