情報検索システムの統計的手法による特徴と精度の分析

佐々木 稔[†] 北 研 二^{††}

本論文では, IREX ワークショップにおける情報検索課題 (IR) の本試験の結果, および, 参加したすべての IR システムについてのアンケートをもとに, 平均適合率, 再現率・適合率曲線を直線回帰させた傾きと切片がシステムに用いられた手法とどのような相関関係をもっているのかを調査し, それぞれの手法がシステムの性能に与える影響の大きさを示した. その結果, 多くの手法について, 再現率 0.0 での適合率の値と適合率の減少量にトレードオフの関係が存在し, 検索システムに用いる手法の選択の難しさが現れた. また, NARRATIVE タグの使用有無により, 同様に相関関係を調査し, NARRATIVE タグの有効性とシステムの性能に与える影響の大きさを示した. その結果, NARRATIVE タグを利用する場合, それに適した有効な手法を選択することが重要であることが分かった.

キーワード: 情報検索, IREX ワークショップ, 性能評価, 適合率, 再現率

Analysis of the Characteristics and the Efficiency of Information Retrieval Systems by Statistical Method

MINORU SASAKI[†] and KENJI KITA^{††}

In evaluating the effectiveness of information retrieval (IR) and extraction system, the most common method is to compare two retrieval methods and decide if one system measurably achieves better results than the other. However, it is difficult for researchers to compare more than two retrieval methods because there are many participants in IR task in IREX workshop. In this paper, we evaluate the characteristics and the effectiveness of the IR systems using a statistical method based on the results of the IR formal run and questionnaires of systems. Comparisons of systems deal with effects on the performance such as indexing, querying and retrieval model. The results confirm the effectiveness of this evaluation method because phrases relates to the performance better than words. There is a trade-off relation between the precision value at 0.0 and decrease rate in many systems and this result indicates the difficulty of the choice of techniques in system. We also evaluate correlations between the efficiency and the characteristics of the systems with both a short and long versions of the topics. A result of this evaluation shows that it is important to select effective methods for the long version of topics.

KeyWords: *Information Retrieval, IREX Workshop, Evaluation, Precision, Recall*

[†] 徳島大学大学院工学研究科, Graduate School of Engineering, University of Tokushima

^{††} 徳島大学工学部, Faculty of Engineering, University of Tokushima

1 はじめに

情報検索の分野は、欧米において過去数十年の間に、英語を中心とした文書を対象に研究が盛んに進められ、高速な文字列検索アルゴリズムや自動索引づけなどに多くの成果が得られた。これらの技術が基礎となり、大規模な文書集合に対する検索技術、新しい評価技術の向上を目的として、TREC (Text REtrieval Conference)¹などのコンテストが開催され、新しい技術の開発やこれまでの技術の改良などが活発に行われている。

日本においても、情報抽出、検索技術に関する研究が盛んに行われ、数多くの優れた日本語情報検索システムが提案されている。このようなシステムを評価するための日本語テストコレクションの整備も進み(木谷 1998)、個々の検索システムを容易に評価できるようになった。さらに、共通のデータベース、プラットフォームにおけるシステム評価の場として、IREX (Information Retrieval and Extraction Exercise) ワークショップ²が開催された。このワークショップには、情報検索(IR)と情報抽出(NE)の各課題に対して数多くのシステムが参加し、全体的な評価を通して様々な議論が行われた。

IREXの目標のひとつとして、共通の基準における各検索システムの評価を基にした問題点の共有と、それによるこの分野の飛躍的な進歩、発展がある(関根, 井佐原 1999)。一般的に、情報検索システムの性能評価をする際には、提案された手法を利用したシステムと利用していないシステムとの比較を行う。比較する際、一つのシステムからみると、参加した数多くのシステムにおける評価結果の違いから、研究の新しい方向性や発展性が発見できる。しかし、IREXでは、数多くのシステムが参加しているため、ふたつのシステム間の比較実験では実験回数が莫大となり、共通点、相違点の整理が複雑になってしまう。また、他の比較手法として、使用されたシステムとは別の基準システムを作り、比較を行う手法も提案されている(Hull 1993)。しかし、その場合、システム間の相違点が多くなり、直接的に何が精度向上の原因であるのかをとらえることが難しくなる。したがって、すべての検索システムを対象として、システムの構成要素を評価すると同時に、全体的なシステムの検索精度を評価するようなシステム指向の評価方法が必要となる。このような全体的な評価は、問題点を発見、解決するための議論を進める上で重要な課題であると考えられ、TRECやIREXでは様々な評価が行われている(Lagergren and Over 1998)(Voorhees 1998)(松尾, 宮本, 森 1999)。

本論文では、IREXにおけるIR課題の本試験の結果、および参加した各システムについての、参加者が回答したアンケート結果を参考にして、IR課題におけるシステムの特徴と精度の関連性を独自の統計的な手法を用いて分析する。これまでは、手法を利用したシステムと利用していないシステムとの実験結果を比較することによって、その手法の有効性が評価されていた。これに対し、我々の提案する評価手法は、数多くのシステムにおける検索結果を基にして、

1 TREC ワークショップホームページ: <http://trec.nist.gov>

2 IREX ワークショップホームページ: <http://cs.nyu.edu/cs/projects/proteus/irex/>

システムに用いられた手法との関連性を客観的な相関係数として表し、検索システムに対し有効な手法を明確にしている。このような検索システムに対し有効な手法を示す評価は、これまで TREC7 においても行われているが、比較に用いられたすべてのシステムで再現率・適合率曲線の違いがほとんど無い条件の下で行われている (Voorhees and Harman 1998)。この条件において、比較に用いたシステムが利用した手法が示されているが、客観的にその手法が有効かどうかの判断は難しい。その点で、我々の評価手法はどのような再現率・適合率曲線に対しても客観的に有効な手法を示すことができる。さらに、我々の評価手法は、検索結果でのランクの上位に、関連のある文書を数多く検索するための有効な手法を示すことができる。この分析においては、IR システムのアンケートの中でシステムの性能に大きく影響する次の 3 点

- 索引づけ、索引構造
- 検索式の生成
- 検索モデル、ランクづけ

に注目して、これらの要素を実現するために用いられた手法が検索精度とどの程度関連があるのかを調査する。

IREX では、検索課題³に、検索要求を簡潔に表現した DESCRIPTION タグと、人間が判断可能な程度の詳細な検索要求の記述をした NARRATIVE タグが用いられている。通常、WWW サイトなどに存在する検索エンジンに入力される索引語の数は 2, 3 語と少ないために、DESCRIPTION タグのみを検索実験に考慮する方が実用的である。しかし、DESCRIPTION タグのみを利用した場合には曖昧さが生じてしまい、人間が可能な限り正確に検索できるという点においては、詳細に書かれている NARRATIVE タグの方が重要な情報であるといえる。実際、TREC などにおいても、このような検索要求の長さに対する精度への影響が議論され、NARRATIVE タグの使用による精度の違いが分析されている (Voorhees and Harman 1997)(Voorhees and Harman 1998)(Hull 1996)。このようなことから、IREX においても、検索式を作成する際の NARRATIVE タグの使用有無により、検索システムに与える影響が変化するものと考えられる。このことを明らかにするため、検索システムにおける NARRATIVE タグの利用有無により short と long に分け、それぞれの平均適合率と関連の高いシステムの特徴を調べる。

再現率・適合率曲線に対し単回帰分析を行い直線として近似した場合、その切片が大きい時、ランクの上位に適合する文書を検索できる確率が高いと考えられる。また、傾きが平行に近い

3 検索課題の例としては、以下のようなものがある。

```
<TOPIC>
<TOPIC-ID>1001</TOPIC-ID>
<DESCRIPTION> 企業合併 </DESCRIPTION>
<NARRATIVE> 記事には企業合併成立の発表が述べられており、その合併に参加する企業の名前が認定できる事。
また、合併企業の分野、目的など具体的内容のいずれかが認定できる事。企業合併は企業併合、企業統合、企業買収も
含む。</NARRATIVE>
</TOPIC>
```

図 1 A 判定のみの再現率・適合率曲線

ほど、システムは再現率の増加とともに起こる適合率の減少を抑えることができると考えられる。そこで、検索結果を平均して得られた再現率・適合率曲線に単回帰分析を行い直線として近似し、その切片と傾きがさまざまな手法のなかでどの手法に関連性が強いのかを調べる。また、同様に、short と long における切片と傾きとの相関が高いシステムの特徴を調べる。これら进行分析することにより、本試験に参加したすべてのシステムで、検索質問を short と long に分けたそれぞれの場合に対して、傾き、切片から総合的に、どの手法と関連性が強いかを考察する。

2 再現率と適合率の関係

IR システムの評価には、一般的に適合率 (Precision) と再現率 (Recall) が使用される (Lewis 1991)(Witten, Moffat, and Bell 1994)。適合率は、システムが検索した文書に対する、検索した正解文書数の割合であり、検索の能力を表している。再現率は、全正解文書数に対するシステムが検索した正解文書数の割合であり、システムがすべての適合する文書のうちの程度実際に検索可能かという検索の幅を表している。再現率と適合率はそれぞれ個別に用いてもシステムの評価を行うことができるが、ランクづけを行う検索システムでは、一般的に再現率・適合率曲線が用いられている。

IREX ワークショップにおける IR 課題は、2 年分の新聞記事から検索課題に書かれた検索要求に関連する文書を検索するもので、予備試験と本試験が行われた。予備試験は課題数が 6 課題あり、評価結果は非公開で、本試験は課題数が 30 課題あり、評価結果は実際の団体名が分か

らないように各団体をシステム ID により表し、公開している。検索結果の判定基準には A, B, C の 3 種類あり、A 判定は記事の主題が検索課題に関連している場合、B 判定は A 判定のように記事の主題には関連性がないが、記事の一部が関連している場合、C 判定は何も関連していない場合という判断基準になっている。その IR の本試験に参加したすべてのシステムにおける、A 判定のみを正解とした検索結果を図 1 に示す。このグラフにおいて、一つの曲線を示す '1103a' などの文字列はシステム ID を表している。このグラフから分かるように、再現率と適合率の関係は大きく分けて 2 種類あると考えられる。一つ目は、適合率が再現率の増加に対し、直線的に減少する関係である。これは、ほとんどのシステムに当てはまる傾向で、特に再現率が 0.0 での適合率の値が高いシステムがこのようなグラフになっている。もう一つはグラフが下に凸の曲線を描くように適合率が減少する関係である。このような曲線は、検索の結果上位にランクされている文書に適合する文書が少ないため、再現率が少ない値において適合率の変動が激しくなっていると考えられる。システムについてのアンケートから、この曲線になる直接的な原因を調査したが、特にシステムに共通して用いられている手法は存在しなかった。

また、多くのシステムが再現率 0.0 における適合率が 0.7 を超えており、ランク上位に適合する文書が検索される確率が高くなっている。再現率 0.0 における適合率の値はランク上位に適合する文書を検索できるのかを表す尺度で、高い値をもつシステムほどユーザの探している情報が検索されていると考えることができる。ランクの 1 位で検索された文書が適合する文書であると、その時点で再現率 0.0 での適合率は 1.0 となる。適合しない文書の場合には適合率は 1.0 にはならず、以降の検索結果のランクに従って適合率に大きな変動が生じる。変動の大きさはあるものの、その後適合する文書が適合しない文書に比べ数多く検索されると、適合率は高い値となる。

しかし、再現率 0.0 での適合率の大きさには関係なく、再現率・適合率曲線のグラフは必ず単調に減少する特徴を持っている。ランクに関係なく適合する文書が連続して検索される、または適合する文書の割合が適合しない文書に比べ大きい場合には、再現率の増加に対する適合率の減少量が少なくなる。ランクの上位にいくらか適合しない文書を検索していたとしても、途中で適合する文書を連続して検索し、適合しない文書の数に比べて多くなることで、適合率の減少が少なくなり、傾きの最大値である 0 に近くなる。これより、適合する文書を効率良く検索するためには、グラフの減少量を少なくする必要がある、これはシステムを評価する上で重要な尺度であると考えられる。

3 評価実験

IREX ワークショップにおける、IR 本試験の結果を公表する方法については、団体名を実名で公表するのではなく、各団体に割り当てられた ID 番号で結果を公表する方法がとられた。このため、検索結果に対してどのような手法が用いられているのか、研究的な内容を対応づけることが難しくなっている。システムの詳細を知る手段として、検索結果と同時に提出した IR システムアンケートがある。このアンケートにより、各システムがどのような手法を利用したかが理解できるようになっている。

IR システムアンケートは、各システムについてどのような手法を用いて本試験の検索結果を出したのかを回答したものであり、主に次の項目がある。

- 索引づけとそのデータ構造
- 検索式の作成
- 検索を実行する環境
- 検索モデル
- その他

これらの項目の回答を集計して、システムが用いた手法の内、主要な 55 個の手法に注目した。これらの手法を平均適合率などの数値と比較するために、手法の使用の有無により数値を割り当てる。本実験では、各システムが用いた手法には 1 を、用いていない手法については 0 を割り当て、数値データに変換した。このときアンケートの中で「はい」、「いいえ」で答えられない質問項目については、システムにただひとつしか用いられていない手法でもひとつの手法として数値を割り当て、システムの違いを明確に定めることにした。このように変換をしたデータに平均適合率を付け足してひとつの行列にして、用いられた手法と平均適合率との相関係数を求めた。相関とは、変数 x , y において一方の変化が他方の変化にある傾向を伴うとき、 x と y の間に相関関係があるという。相関には次の 3 種類がある。

正の相関：一方の値が大きくなると、他方の値も大きくなる。

負の相関：一方の値が小さくなると、他方の値も小さくなる。

無相関：2 つの値に明白な関係がみられない。

この相関を数値的に表したものに相関係数があり、2 つの変数の相互関係の程度を表したものである。本実験では、相関の有無を明確にするために、相関係数の絶対値が 0.5 を超える手法は相関が認められるとして、システムが用いた手法に対して評価を行った。

3.1 平均適合率とシステムの関連

平均適合率とシステムとの相関係数を求めた結果、相関係数の高かった主なシステムの特徴を表 1 に示す。相関係数の高い手法は、全体的に正の相関を持っているが、すべての手法に対し、

表 1 平均適合率と相関の高い主なシステムの特徴

システムの特徴	平均	相関係数
LSI	0.04545	-0.49600
IDF	0.86364	0.49341
レレバンスフィードバック	0.22727	0.45499
名詞	0.45455	-0.42376
フレーズ	0.31818	0.41665
NEG タグ	0.36364	0.40655
ロバートソン法	0.09091	0.40427
文書の長さ	0.45455	0.38297
名詞以外の品詞	0.40909	-0.38053
シソーラス	0.13636	0.33728
DESCRIPTOIN タグ	0.90909	-0.33728
文字列形態素インデクス	0.09091	0.32822
BM25	0.09091	0.32822

相関係数の絶対値が 0.5 を下回っている．表 1において，相関係数の絶対値の高い LSI(Latent Semantic Indexing) や IDF(Inverse Document Frequency) については，若干の相関は見られるものの，0.5 に満たしていないため，平均適合率との間に明確な相関を認めることができなかった．

また，この表 1を見ると，DESCRIPTOIN タグから索引語を抽出する手法と NEG タグなどのように NARRATIVE タグを利用した手法が混在していることが分かる．実際に，NARRATIVE タグ自体を利用することについては，平均適合率との相関係数が -0.03864 と低く，NARRATIVE タグは他の 2 つのタグに比べ，適合率との関連性が少ない結果となった．これは，NARRATIVE タグの中には比較的長い文章が存在し，検索に重要な索引語が存在するのと同時に，一般的に広い分野で使われる索引語も比較的多く存在しているため，このタグを使う際には注意が必要であると考えられる．このようなことから，IREX ワークショップに参加したシステムを検索課題を簡潔に表現した DESCRIPTOIN タグのみを用いたシステムと比較的長い文章が存在し，NEG タグ⁴が存在する NARRATIVE タグを同時に用いたシステムに分けて評価を行う．これにより，相関の認められる手法がより顕著に現れ，より有効な評価をすることができると考えられる．

4 NARRATIVE タグ中において，「～を除く」などの否定的な表現を示すもの

表 2 short の平均適合率と相関の高い主なシステムの特徴

システムの特徴	平均	相関係数
IDF	0.88888	0.64416
LSI	0.11111	-0.64416
フレーズ	0.55555	0.61796
レlevanceフィードバック	0.33333	0.57326
確率と情報量	0.33333	0.56095
名詞	0.66667	-0.54080
名詞以外の品詞	0.66667	-0.54080
構文解析	0.22222	-0.43654
NEG タグ	0.22222	0.42834
シソーラス	0.22222	0.42834
単語	0.77778	-0.42834

表 3 long の平均適合率と相関の高い主なシステムの特徴

システムの特徴	平均	相関係数
固有名詞	0.15385	0.74158
ロバートソン法	0.15385	0.74158
構文的な手がかり	0.15385	0.74158
文書の長さ	0.46154	0.68955
NEG タグ	0.46154	0.48372
名詞	0.30769	-0.41387
面の情報	0.15385	-0.41384
IDF	0.84615	0.39269
単語	0.61538	0.37213
形態素解析	0.15385	-0.36284

3.2 short および long での平均適合率とシステムの関連

DESCRIPTION タグのみを用い、NARRATIVE タグを利用しない場合を short、検索課題をすべて使用した場合を long として、それぞれを用いたシステムにおける平均適合率と利用した手法との関連性を比較する。そのために、short, long それぞれにおける平均適合率とシステムが用いた手法との相関係数を求め、NARRATIVE タグを利用しない場合にはどのような手法が有効であるか、また、NARRATIVE タグを利用した場合についても同様に有効な手法にどのようなものがあるかを調査する。その結果、short を用いたシステム、long を用いたシステムの平均適合率との相関係数が高かった主なシステムの特徴を、それぞれ表 2 と表 3 に示す。

表 2 と表 3 から共通して言えることは、short と long に分ける前の表 1 と比較すると相関係数

が全体的に高くなっているということである。すなわち、short と long に分けることによって、それぞれの検索システムに対して、平均適合率により関係深いシステムの特徴が顕著になっている。

short で平均適合率と相関が最も高かったのは、LSI と IDF である。LSI は負の相関を持ち、検索精度を下げる傾向があるため、今回の本試験での結果においては、LSI の利用や索引づけに改良が必要であった。平均適合率と直接関連のある手法に、文書の重みづけとしては IDF、索引語には意味を限定しやすいフレーズが、共に正の相関を持ち検索精度を上げる傾向がある。フレーズを用いるとき、索引語が数多く存在しない場合があるため、レレバンスフィードバックを用いて検索式を拡張させることで、性能の良いシステムが構築できると考えられる。名詞や名詞以外という品詞情報を用いて索引づけを行う方法は、どちらも負の相関が高くなった。これは、元々索引語の数などのように情報量が少ない上に、一般的な単語まで取り出していたために、検索精度が下がる傾向があると考えられる。

long で平均適合率と相関が最も高かったのは、固有名詞である。これは、NARRATIVE タグにある文章が比較的長いめ、索引づけの手法の中でも特定分野にしか出現しない語を抽出できる固有名詞の相関が高くなったと考えられる。また、ロバートソン法や構文的な手がかり、文書の長さといった文書そのものの違いや特徴を明確にする重みづけの手法も相関が高かった。short では平均適合率に関連性があった IDF は、long の場合、これらの手法より相関が低かった。このことより、索引語に対する重みづけ手法は関連性が低いと考えられる。

表 2に見られる手法と表 1に見られる手法を比較すると、同じような手法が全体に見受けられた。すなわち、short の場合は、一般的な検索システムの平均適合率に関連性のある手法がそのまま関連性があるということが出来る。long の場合は、表 3に示した相関と全体の結果とを比較すると、全体的に正の相関を持つものが多かったが、表 1では見られなかった手法が多く現れた。

これらの結果から、NARRATIVE タグの使用有無によって検索システムの平均適合率に関連性のある手法が変化することが分かった。したがって、NARRATIVE タグを用いる際には、平均適合率に関連性のある、適した手法を選択する必要があると考えられる。

3.3 回帰式とシステムの関連

検索結果を平均して得られた、再現率と適合率の関係のデータを単回帰分析を用いて直線近似を行う。計算の結果、傾き(回帰係数)や切片(定数項)に対して相関の高かった手法を、それぞれ表 4と表 5に示す。直線回帰を行ったときの決定係数は最大で 0.99446、最小で 0.69193 となり、22 個のシステムにおける平均が 0.942 と高い値となり、再現率・適合率曲線を直線で近似することが妥当であることを表している。

傾きと最も相関係数の高いものは文書の長さで、相関係数が -0.60396 と相関の認められる

表 4 回帰係数と相関の高い主なシステムの特徴

システムの特徴	平均	相関係数
文書の長さ	0.45455	-0.60396
構文解析	0.04545	-0.49385
IDF	0.86364	-0.44741
固有名詞 (索引)	0.22727	-0.43614
固有名詞 (検索式)	0.22727	-0.36891
単語	0.68182	-0.30771
面の情報	0.09091	0.30267
LSI	0.04545	0.30140

表 5 回帰直線の定数項と相関の高い主なシステムの特徴

システムの特徴	平均	相関係数
IDF	0.86364	0.60051
文書の長さ	0.45455	0.55917
LSI	0.04545	-0.55286
ロバートソン法	0.09091	0.44812
NEG タグ	0.36364	0.42826
フレーズ	0.31818	0.39632
固有名詞 (検索式)	0.22727	0.39059
レlevanceフィードバック	0.22727	0.37556
面の情報	0.09091	-0.36234
照合文字列長	0.33400	-0.34285

数値で、傾きを下げる傾向がある。これは、検索した文書数が増えるにつれて、適合する文書を検索する割合が少なくなることを表している。特に、傾き、切片共に関連が大きい IDF と文書の長さは重みづけ手法であり、文書間での違いや特徴を明確にしているため、検索システムの構築において、重みづけ手法が特に重要であると考えることができる。

切片と最も相関係数の高いものは IDF で、次いで、文書の長さ、LSI となっている。全体的に正の相関を持つ手法が多く、切片を上げる傾向がある。しかし、これらの手法には、傾きにおいて負の相関の高い手法と共通するものが多く存在している。これより、これらの手法に対して、切片の大きさと傾きの大きさにはトレードオフの関係が存在していることが分かる。

表 6 short の回帰係数と相関の高い主なシステムの特徴

システムの特徴	平均	相関係数
シソーラス	0.12500	0.75160
LSI	0.12500	0.75160
自動検索質問拡張	0.12500	0.75160
IDF	0.75000	-0.74788
フレーズ	0.37500	-0.57108
文字列形態素インデクス	0.25000	-0.51490
確率と情報量	0.25000	-0.51490
BM25	0.25000	-0.51490

表 7 short の回帰直線の定数項と相関の高い主なシステムの特徴

システムの特徴	平均	相関係数
IDF	0.75000	0.60051
シソーラス	0.12500	0.55917
LSI	0.12500	-0.55286
自動検索質問拡張	0.12500	0.44812
文字列形態素インデクス	0.25000	0.42826
確率と情報量	0.25000	0.39632
BM25	0.25000	0.39059

3.4 short および long での回帰式とシステムの関連

NARRATIVE タグの有効性、有効な利用方法について考えるため、3.2節と同様に short と long に分割し、評価を行った。傾きや切片に対して相関の高かった手法を、short の場合をそれぞれ表 6、表 7に、long の場合をそれぞれ表 8、表 9に示す。

short における関連性

short の場合、傾きと最も相関係数の高い手法は、シソーラス、LSI、自動検索質問拡張で、これら 3 つは等しい値で、正の高い相関を持つ。しかし、全体的には負の相関を持つ手法が多く、傾きを下げる傾向がある。short では情報が少ないために、索引づけされた語をあらかじめ準備した知識集合であるシソーラスで拡張することにより、傾きを上げる高い正の相関が得られたものと考えられる。さらに、LSI、フレーズ、自動検索質問拡張、文字列形態素インデクスもシソーラスと同様に索引づけの手法で、相関係数は比較的高い値になっている。このことから、検索に有効な索引語の選択や LSI による意味的な表現形式を得る手法が、傾きに深く関わっているといえる。

また、切片と最も相関係数の高い手法は IDF で、次いで、シソーラス、LSI となっている。

表 8 long の回帰係数と相関の高い主なシステムの特徴

システムの特徴	平均	相関係数
ベクトル	0.14285	-0.60621
文書の長さ	0.14285	-0.60621
索引語の長さ	0.14285	0.59558
IDF	0.42857	-0.58837
フレーズ	0.28571	-0.47094
インデクス (ベクトル)	0.92857	0.38171
TF	0.07142	-0.38171
構文的な手がかり	0.07142	-0.38171

表 9 long の回帰直線の定数項と相関の高い主なシステムの特徴

システムの特徴	平均	相関係数
フレーズ	0.28571	0.75913
ベクトル	0.14285	0.62619
文書の長さ	0.14285	0.62619
索引語の長さ	0.14285	-0.60014
NEG タグ	0.57142	0.52856
IDF	0.42857	0.49117
名詞	0.28571	-0.46559
n-gram	0.28571	0.36211

こちらは、全体的に正の相関を持つものが多く、切片を上げる傾向がある。傾きとの相関の高い手法と共通するものが多く、切片でも索引づけの手法が深く関連しているといえる。シソーラスや自動検索質問拡張は傾き、切片ともに正の相関係数を持ち、short の場合、シソーラスが切片と傾きに最も関連の深い手法だと考えられる。しかし、LSI、IDF などは、傾きでは正の相関、切片では負の相関を持っているために、ここでも傾きと切片の間にトレードオフの関係が存在することが分かる。

long における関連性

long の場合、傾きと最も相関係数の高い手法は、ベクトル、文書の長さで、次いで索引語の長さとなる。short と同様に、全体的に負の相関を持つ手法が多く、ベクトルなどは傾きを下げた傾向がある。また、検索要求全体における傾き、切片と同様に、文書の長さ、索引語の長さ、IDF などの重みづけ手法が深く関わっていることが分かる。さらに、ベクトルは文書または検索要求全体を用いた手法であるため、long を扱う際の索引語の増加に従って、関連性がより顕著に現れたと考えられる。

切片と最も相関係数の高い手法はフレーズで、次いで、ベクトルと文書の長さが高い相関を持っている。short 同様、正の相関を持つものが多く、切片を上げる傾向がみられるが、これらのほとんどは重みづけの手法である。しかし、これらの手法は、傾きと負の相関の高い手法と多数共通しているため、long においてもトレードオフの関係が存在していることが明らかになった。また、検索要求の長い long の場合、正の相関が高い NEG タグは、傾きとの相関係数が低く、切片との相関係数は高い。したがって、NEG タグの利用は、ランクの上位に適合する文書を検索する有効な手段であると考えることができる。

4 まとめ

本論文では、IREX ワークショップにおける IR の本試験の結果、および、参加したすべての IR システムについてのアンケートを基に、平均適合率、再現率・適合率曲線を直線回帰させた傾きと切片が IR システムに用いられた手法とどのような相関関係をもっているのかを調査し、それぞれの手法がシステムの性能に与える影響の大きさを示した。その結果、名詞や名詞以外の品詞単語を用いる以上にフレーズが性能向上に関係あり、複数の単語を組み合わせることで意味が限定され、精度に良い影響を与えることを確認することができた。また、再現率・適合率曲線の切片と傾きにトレードオフの関係が多く手法に見られ、検索システムに用いる手法の選択の難しさが現れる結果となった。

さらに、NARRATIVE タグの使用有無により short と long に分け、平均適合率との相関関係、また、再現率・適合率曲線を直線回帰させた傾きと切片にそれぞれどのような相関関係があるかを調査し、システムの性能に与える影響の大きさを示した。その結果、分ける前と比較して、全体的に相関が高くなり、関連が大きいシステムの特徴が顕著に現れるようになった。また、NARRATIVE タグを利用する場合、それに適した有効な手法を選択することが重要であることが分かった。short では、分ける前と比較して、索引語そのものに対する索引づけの手法への関連性がより顕著に現れた。その中でも、シソーラスや自動検索質問拡張は検索性能を上げる関連の深い手法であった。一方、long では文書全体に対する重みづけの手法が性能向上に関係があった。形態素解析やそれにより抽出される名詞単語を用いるよりも、フレーズや固有名詞などを用いる方が性能向上に関連が深く、より要求する意味が明確になっている。このことは形態素解析により索引語を選択する難しさを表現し、形態素から索引語をより有効的に選択する手法が必要であることを示している。

最後に、情報検索システムの用いる手法の選択の難しさを克服するため、本論文におけるデータが今後の研究開発で広く利用され、情報検索システムの分野の進歩と発展につながることを期待したい。

謝辞

本論文をまとめる機会を与えてくださり、有用なデータを提供して下さった IREX 実行委員の方々、及び、IREX-IR の本試験に参加した方々と判定者の方々に心から感謝したい。

参考文献

- Hull, D. (1993). "Using Statistical Testing in the Evaluation of Retrieval Performance." In *Proceedings of ACM SIGIR Conference*, pp. 329–338.
- Hull, D. (1996). "Stemming Algorithms - A Case Study for Detailed Evaluation." *Journal of the American Society for Information Science*, **47**(1), 70–84.
- 木谷強ほか (1998). "日本語情報検索システム評価用テストコレクション BMIR-J2." データベースシステム研究会, 情報処理学会.
- Lagergren, E. and Over, P. (1998). "Comparing Interactive Information Retrieval Systems Across Sites: the TREC-6 Interactive Track Matrix Experiment." In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 164–172.
- Lewis, D. D. (1991). "Evaluating Text Categorization." In *Proceedings of Speech and Natural Language Workshop*, pp. 312–318.
- 松尾衛, 宮本昌幸, 森辰則 (1999). "情報検索タスクにおける人間による正解判定についての分析." In *Proceedings of the IREX workshop*, pp. 15–22.
- 西野文人 (1996). "日本語テキスト分類における特徴素抽出." 自然言語処理研究会, 情報処理学会.
- 関根聡, 井佐原均 (1999). "IREX プロジェクト概要." In *Proceedings of the IREX workshop*, pp. 1–5.
- Voorhees, E. M. (1998). "Variations in Relevance Judgements in the Measurement of Retrieval Effectiveness." In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 315–323.
- Voorhees, E. M. and Harman, D. (1997). "Overview of the Sixth Text REtrieval Conference (TREC-6)." In *Voorhees and Harman (Eds.), Proceedings of The Sixth Text REtrieval Conference (TREC-6) National Institute of Standards and Technology Special Publication*, pp. 1–24.
- Voorhees, E. M. and Harman, D. (1998). "Overview of the Seventh Text REtrieval Conference (TREC-7)." In *Voorhees and Harman (Eds.), Proceedings of The Seventh Text REtrieval Conference (TREC-7) National Institute of Standards and Technology Special*

Publication, pp. 1–24.

Witten, I. H., Moffat, A., and Bell, T. C. (1994). *Managing Gigabytes: Compressing and Indexing Documents and Images*. Van Nostrand Reinhold, New York.

略歴

佐々木 稔: 1973 年生. 1996 年徳島大学工学部知能情報工学科卒業. 1998 年徳島大学大学院博士前期課程修了. 同年, 徳島大学大学院博士後期課程入学, 現在に至る. 機械学習, 情報検索等の研究に従事. 情報処理学会会員.

北 研二: 1957 年生. 1981 年早稲田大学理工学部数学科卒業. 1983 年から 1992 年まで沖電気工業 (株) 勤務. この間, 1987 年から 1992 年まで ATR 自動翻訳電話研究所に出向. 1992 年 9 月から徳島大学工学部勤務. 現在, 同教授. 工学博士. 確率・統計的自然言語処理, 情報検索等の研究に従事. 情報処理学会, 電子情報通信学会, 日本音響学会, 日本言語学会, 計量国語学会, ACL 各会員.

(2000 年 5 月 16 日 受付)

(2000 年 7 月 12 日 再受付)

(2000 年 10 月 10 日 採録)