DB2の全文検索エンジンにおける セレクティビティの実装

山本誠 Makoto Yamamoto
是津耕司 Kohji Zettsu
日本アイ・ビー・エム株式会社 大和ソフトウェア開発研究所情報マネージメント技術第一開発

概要:

データベースにおける検索順の最適化のために各項目に対する検索結果の大きさを見積もる必要がある。 本論文では、全文検索エンジンＧＴＲに検索結果の文書数を見積もる機能「セレクティビティ」で使用されている手法と有効性について論じる。

はじめに

紙メディアに記録されていた情報のデジタル情報への置き換えが進み、デジタル情報に含まれる文書量も増えてきている。さらに、インターネット技術の発達に伴いネットワーク上に蓄えられるＷｅｂサーバなどに蓄えられる文書量も増加している。ＤＢ２ではさまざまなデータを扱うことができるように各種のエクステンダーが用意されており、テキストデータを取り扱うためのエクステンダーも用意されている。ＧＴＲはＤＢ２においてテキストエクステンダーやネットサーチエクステンダー[2]の一部として組み込まれ全文検索を行っているコンポーネントである。

ＤＢ２におけるＳＱＬのサーチパスの最適化には事前に照会に含まれる列を検索するために必要なコストを見積もる必要がある。このためユーザー定義関数では照会中にセレクティビティ節があり[2]予想される結果セットの大きさを指定することができる。ＧＴＲにおける結果セットの大きさは検索結果の文書数に対応し、検索語によって大きく変動する。このため、検索結果の文書数を見積もる機能「セレクティビティ」をＧＴＲに新たに実装した。

本論文ではセレクティビティにおいて使用されている手法および、性能の評価を行い有用性について検証する。

1. DB2におけるサーチパスの最適化

ＤＢ２において照会に複数の列が含まれている場合には検索順の最適化を行い照会にかかる時間を短縮している。例えば、単純な照会

```
SELECT *  FROM  EMPLOYEE WHERE
NAME=...  AND AGE=...
```

の場合 について考える。 行数が１０００のデータベースに上記のような照会を行う場合を考える。列ＮＡＭＥとＡＧＥを個別に検索した検索の結果、それぞれの検索の結果セットの大きさが２と１００あり、最終結果として１つの行が選択されるとする。このとき、ＮＡＭＥに対する検索の結果からＡＧＥに対する条件に合うものを検索した場合には、ＡＧＥに対する検索結果からＮＡＭＥに対する条件で検索を行った場合と比較して、結果は同じであるが中間結果の大きさが大きく異なるため、検索のコストが検索対象の大きさに比例すると考えると、検索にかかるコストは少なくなる（図１）。

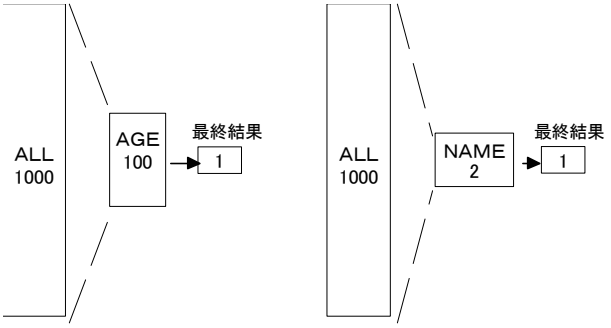

図１：検索順の最適化の概念図

さらに、文書に対する全文検索を含むような複雑な照会

```
SELECT * FROM  WEBINDEX WHERE
TITLE="estimate * method" and SERVER=...
```

のように、列ＴＩＴＬＥに対する全文検索が含まれている場合であっても、それぞれの列に対するの検索結果の大きさを予め取得できるならば、サーチパスに効果的な最適化を行うことができる。 このように各列に対する結果セットの大きさはサーチパスの最適化において非常に重要である。

しかし、これまでは全文検索にかかるコストを事前に評価する機能が存在しなかった。このため前述のような最適化を行うことができず効率的な検索を行うことが困難であった。このような困難を克服するために、検索にかかるコストの事前評価を行うことができる機能セレクティビティを実装する必要があった。

2.セレクティビティの実装

ＧＴＲを用いた全文検索では検索条件に合った文書と、文書中における検索語の位置情報を結果として返している。 また検索の過程でＡＮＤやＯＲなどの論理操作も行っている。ＧＴＲにおける全文検索の過程では

1. 検索語からそれぞれの語を切り出す
2. 各語をNグラムの手法により検索キーに分割
3. それぞれの検索キーの情報を検索する
4. 検索キーを含む文書、位置情報を取得
5. 検索語に対する条件により論理操作や、ランク付けを行う

ということを行っている。検索語を含む文書数の情報は上記の過程の3の段階で得られる。しかし、文書中の位置の情報はサーチパスの最適化には使用しないためセレクティビティではそれ以降の過程は実行せず、代わりに検索語に対する独自の簡略化された論理操作を行っている。論理操作の簡略化の例として

```
search  * term
```

という検索語を検索する場合を考える。ここで、"*"は論理操作のＡＮＤを表す。索引に含まれるそれぞれの単語を含む文書数をN(search),N(term)で表し、

N(search)>N(term)であるとすると。実際の検索の結果はN(term)よりも小さくなる。このことからセレクティビティでは上記の検索結果の見積もりとして N(term)を返す（図２斜線部）。

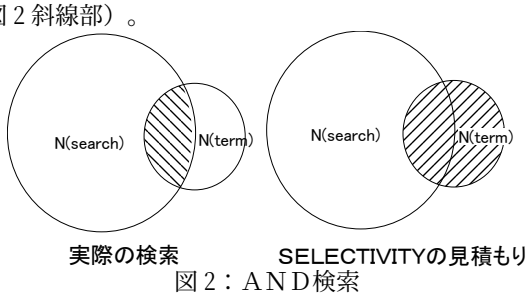

図２：ＡＮＤ検索

同様にＯＲの検索条件では

min(N(search)+N(term),Nall)

を返す。ここで min は最小値を返す関数、Nall は全文書数である。つまり、セレクティビティでは通常、実際の検索結果よりも検索結果に含まれる文書数は多く見積もられるようになっている。しかし、文書中の位置の特定を行わないことや論理操作を簡略化によって高速な応答を実現する。

3. セレクティビティの性能評価

セレクティビティの評価をおこなうため、同一の索引に対して通常の検索とセレクティビティを用い、所要時間と検索結果の文書数の比較を行った。通常の検索は詳細なヒット情報を受け渡さないように指定するなど検索時間を最小化するようにして測定を行った。今回の測定では英語のＷｅｂページで文書数が数約 7,000,000 文書、索引の大きさが１２ＧＢの索引を用いた。測定環境は

IBM pSeries 610,1-Way Power3 - II 450 MHz

である

３．１　実行時間の評価

セレクティビティは通常の検索より短時間で実行することが必要である。評価として、通常の全文検索とセレクティビティの実行時間を測定した。

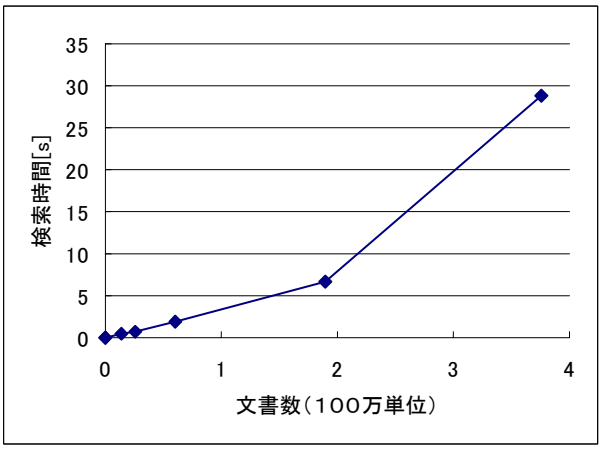

図３：通常の検索にかかる時間

上図３からもわかるように、通常の全文検索においては検索語が多くの文書に含まれる場合には検索にかかる時間が増加する。

一方セレクティビティにかかる時間は、通常の検索と同じ検索語に対して常に 0.01 秒以下で変化しなかった。実際に検索を行うために必要な時間と比較すると、見積もりに必要な時間は非常に小さく、サーチパスの最適化に使用するための機能として十分な速度である。

３．２　実際の文書数と見積もられた文書数の比較

同一の索引を用いて、同一の検索語に対して実際の検索を行った場合にヒットした文書数とセレクティビティによって見積もられた文書数を比較した。

検索語	通常の検索	SELECTIVITY
IBM	3750024	3750024
Search	1903083	1903083
IBM − Search (NOT)	2810413	3750024
IBM ＊ Search (AND)	939611	1903083

表１：実際の検索とセレクティビティの見積もりによる文書数の比較

このように、単一の語のみを含む検索語に対する検索では実際の検索と見積もりは同じ値を返すが、複数の語から構成される検索語に対しては、実際の検索よりも多く見積もられる。

今回の検証では、実際の文章数の数倍程度で実際の文書数を見積もることができており、サーチパスの最適化には十分な精度である。

4. まとめ

全文検索エンジンＧＴＲに新たにセレクティビティの機能を追加した。その結果、ＤＢ２におけるサーチパスの最適化に必要な機能を提供することが可能になった。

セレクティビティに求められる能力

1. 検索結果の文書数を見積もる
2. 実際の検索よりもはるかに短時間で実行可能

の２点について検証し、必要な能力があることを確認した。

現在ＤＢ２にはセレクティビティの機能をもったＧＴＲは組み込まれていないため本稿でのふれたようなサーチパスの最適化は行われていないが、セレクティビティ機能をつかった最適化が行われるようになればテキストエクステンダーやネットサーチエクステンダーの性能向上が期待される。

謝辞

セレクティビティの開発をするにあたり、様々なご協力を頂きました GTR チームのみなさまに深く感謝いたします。

参考文献

[1] DB2 ネットサーチエクステンダーの解説
http://www.ibm.com/jp/software/data/db2/extenders/netsearch.html
[2] DB2 におけるセレクティビティの解説
http://www.ibm.com/jp/software/data/developer/library/manual/db2online/db2s0/ch2srch.html