デジタル・ドキュメント 15－2
（1998. 11. 27）

分散型大規模文書検索システムに関する一検討
A Study on a Distributed Full-Text Retrieval System

森 大二郎　　大森 信行　　田中 一男
Daijiro MORI　　Nobuyuki OHMORI　　Kazuo TANAKA
NTT ヒューマンインタフェース研究所
NTT Human Interface Laboratories

概要

増大し続けるWWW(World Wide Web)の情報空間から文書情報を網羅的に収集し、不特定多数の利用者に対して検索機能を提供する文書検索システムに関する検討内容を報告する。WWW情報の特性と動向から、このようなシステムにおいては、スケーラビリティと情報の速報性が特に重要な要件であると考え、PCクラスタによる分散サーバ構成を用いることとした。実験の結果から、このシステムが高いスケーラビリティを備えていることを確認することができた。

This paper describes about a study on a full-text retrieval system which allows people search the whole documents consisting of the World Wide Web. In order to keep the system useful, it must provide the high scalability and be able to update its index frequently enough so as to keep the contents fresh. We designed the system as a distributed servers based on a cluster of PCs. Through the tests on a small experimental system, we confirmed the high scalability of the system.

1 はじめに

社会の情報化が進行すると共に、我々をとりまく情報環境の有り様も大きな変貌を遂げつつある。我々は従来マスメディアを介して流通する膨大な情報に晒されながら、しばしばごく非効率的なやり方で必要な情報を取り出し、活用していた。インターネットに代表される情報通信の普及は、こうした一方向的な情報授受の関係を一変させ、流通する情報の容量は勿論、その内容や品質の多様さを一気に拡大した。こうした動向は、ともすれば情報過多の問題を更に助長する要因ともなり得るが、一方、ここで流通する情報の多くが計算機に可読なデジタルドキュメントの形態となっているという事実は、これらの情報を自動的に加工し、目的に合致する情報を能動的に収集し、有用性に富む知識を提供することのできる新たな情報環境に発展する可能性を示唆している。

全文検索技術は、こうした新たな情報環境の端緒を開く道具の一つであり、既にWWW(World Wide Web)の文書情報を検索対象とする検索サービスが多くの人々に活用されている。本稿では、このような大規模で、かつ、その内容が流動的に変化する文書集合を検索対象とする文書検索システムの検討内容について報告する。

2 背景

我々は、従来から開発が進められてきた全文検索技術を応用することによって、インターネット上に存在する文書情報を網羅的に収集し、これを検索可能とする、大規模文書検索システムを構築することを検討している。このようなごく大規模な文書検索システムは、原理的には従来技術を演繹することによって実現可能であるが、より多くの人々が実用的に利用できるシステムを構築するためには、さらにいくつかの条件を満足する必要があると考えられる。

2.1 スケーラビリティ

インターネットを流通する文書は現時点でも膨大な量に及ぶが、その量はさらに今もなお急速に増大し続けている。WWW (World Wide Web)のリソースを例にとってみれば、1993年以来、WWWのサーバ数は半年で2倍以上のペースで増加し続けている[1]。このようなデータ規模の成長率は、従来の文書検索システムが扱って来た対象の範囲を大きく越えている。

したがって、こうした情報空間を対象とする情報検索システムにおいては、従前のシステムが想定してきた範囲を越えて、管理するデータの規模を柔軟かつ容易に拡張できるスケーラビリティを実現することが必須の条件となる。

2.2 情報の速報性

当初WWWは学術研究における文献や資料を相互に関係付けながら管理することを意図して開発され、どちらかと言えば変動の少ない静的な情報を保管する書庫のような利用形態が想定されていた。しかし今やニュースやトレンドデータのような速報性が重要な意味を持つ情報や、BBSや投票システムのような相方向の情報更新を伴う情報などの様々なコンテンツがこのメディアに溢れており、ストック型・フロー型という分類の枠組を越えて利用形態が多様化している。

また、文書構造の明示的な表現と操作の体系化を促進するRDF[2]やDOM[3]のような規格が普及すれば、電子化文書の用途はさらに拡大し、即時性を必要とする情報が大量に流通するよ

うになると考えられる。

従来、情報検索システムはDBMSのようなトランザクションシステムと較べて、データの更新頻度が少なく、また更新処理には程度の差はあれ一定のオーバヘッドを伴うのが普通であった。しかし、多数の人にとって利便性の高い情報を提供する情報検索システムにおいては、このような電子化文書の用途の変容に追従し、情報の更新をスムーズに行い、常に鮮度の高い情報を供給できることが必要となる。

3 全文検索エンジン

前章に述べた要求条件の実現手段について説明する前に、本システムの主要な構成要素である全文検索エンジンの特徴について概説する。この全文検索エンジンは、WWWの情報検索検索システムを構成する上で今や前提ともなっている3つの要求条件についてすぐれた特徴を示している。

1. 高速な検索性能
2. 容易な検索指示
3. 高精度な検索結果

以下にこれらの特徴を実現する方式について説明する。

3.1 単語インデクス方式

本エンジンでは、検索対象となる文書を形態素解析により単語に分割し、この中に含まれる全ての自立語をキーとする単語インデクスを作成する。単語インデクスを用いることにより、単語文字列をキーとして、この単語を含む文書のリストを高速に取得することができる。インデクスは単語文字列を構成する各文をキーとするトライ(trie)構造を用いて実装している。トライは、キーとなる文字列を先頭の文字から順番に取り出し、各文字毎に分岐する木構造をなす。トライにおける探索時間はキーの文字列長に比例し、キーの数による影響を受けないためハッシュ等のアルゴリズムに較べて高速である場合が多く、また、挿入・削除の操作が高速に行えるという利点を持っているため、文字列のインデクシングに好適であるとされている[5]。

3.2 自由文による検索

本エンジンでは、検索条件となる文字列として任意の語句あるいは文を受理する。検索文字列は検索対象となる文書と同様に形態素解析処理により単語に分割される。それぞれの単語を含む文書集合を単語インデクスより取得すると、これらの集合に対する明示的なBoolean演算指示が与えられなくとも、後述のTF×IDF法に基づいて重み付けを施した上でそれぞれの結果を結合し、高精度な検索結果を返すことができる。

3.3 検索結果の順位付け

検索結果が複数得られた場合には、与えられた検索文字列に対する適合度を計算し、適合度の高い順に結果を提示することにより、利用者が意図した検索結果を選び出す過程を支援する。適合度はTF×IDF法[4]に基づいて計算する。全文書集合を通して出現頻度が低い単語は特徴

的であり、その重要度が高いという原理に基づき、各単語の重みを計算する。単語 i の重み (idf_i) は以下の式により算出する。

$$idf_i = \log_2 \frac{N}{n_i}$$

ここで、N は全文書数を、n_i は全文書の中で単語 i を含む文書数を示す。

さらに、検索条件に与えられた単語をより多く含む文書ほど適合度が高いという原理に基づき、検索結果である各文書の適合度を計算する。文書 j の適合度 ($score_j$) は以下の式により算出する。

$$score_j = \sum_{i=1}^{k} \frac{\log_2 (tf_{ij} + 1) \times idf_i}{\log_{10} len_j + 1}$$

ここで、tf_{ij} は、文書 j に単語 i が出現する回数を、len_j は文書 j を構成する単語の総数を、k は検索文字列に含まれる単語の数を示す。

この適合度 ($score$) の大きなものから順に検索結果を提示する。

4 分散化へのアプローチ

前章に述べた検索エンジンでは、数 G バイト程度の規模の文書集合において、実用的な検索速度を実現している。しかし、検索エンジンが動作するハードウェアが実装するメモリ容量や OS のサポートするファイルサイズの上限により、数 G バイトを越えるサイズの文書を十分な速度性能を維持しつつ管理することが困難であった。我々は文書集合を複数の部分集合にパーティショニングし、複数の検索サーバで各部分集合を分散管理するアプローチを取ることによって、この問題の克服を試みた。我々のアプローチの特徴を以降に示す。

4.1 PC クラスタの採用

従来、並列処理においては、MPP(Massive Parallel Processers) と呼ばれる、個々のプロセッサの能力やストレージ容量が比較的小さいノードを大量に集め、高速な通信手段で接続する構成が使われて来た。しかし、近年は比較的高速なプロセッサとローカルディスクを備えたノードを比較的低速な通信手段によって結合するクラスタコンピューティングが注目を浴びている。PC クラスタは大規模な市場を持つ PC の安価な部品を使用してクラスタコンピューティングを実現する方法である。構築やノードの追加に要するコストが低いことは、スケーラビリティを高める要素の一つとして評価できる。

我々はまた、WWW における情報検索においては、大量の検索結果が得られたとしても、実際に利用者に閲覧されるのは、適合度が高いごく一部のデータに限られているという事実に着目し、一度の検索処理について分散サーバ間で授受されるデータサイズは平均的にはごく小さくて済むと判断し、低速だが廉価な通信手段を備えた PC クラスタを採用することにした。

4.2 データパーティショニング

分散サーバ上にデータを配置する方法についてはいくつかの選択肢が考えられる。例えば、検索キーとなる単語の集合を複数のサブセットに分割する方法や、検索対象となる文書の集合を複数のサブセットに分割する方法等である。

単語によってサーバを分割した場合、与えられた検索文字列から検索処理を行うべき検索サーバを容易に特定することができる反面、検索文字列として複数の単語が与えられた際に、複数のサーバ間で全ての検索結果の情報を交換しなければTF×IDF法に基づく適合度によって結果をソートすることが出来ない場合がある。

文書によってサーバを分割した場合、基本的には全ての検索要求を全ての検索サーバに配布しその応答を待たねばならないが、各検索結果の適合度はそれぞれの検索サーバで独立に計算することが可能となる。我々は、通信量を低く抑えることを重視し、文書によってサーバを分割することにした。

図1にシステムの概要を示す。システムは、一つ以上のフロントエンドサーバと、複数の検索サーバ、及び情報更新サーバから構成される。この構成は物理的な構成と一致する必要は必ずしもなく、例えば、フロントエンドサーバと検索サーバ、検索サーバと情報更新サーバ等を同一のハードウェア上に構成することができる。しかし、数Gバイトを越える大規模な文書集合を対象とする場合には、複数の物理的に分散した検索サーバが必要であり、それぞれのハードウェアが適当なサイズのローカルディスクと主メモリとを備えていなければならない。

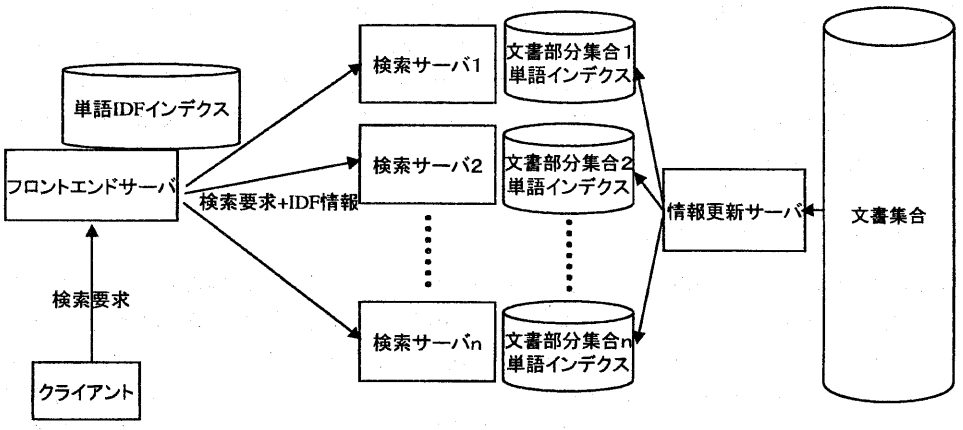

図1 システム構成

検索対象となる文書集合は、情報更新サーバにおいて検索サーバの数に一致する複数のグループに分割され、それぞれのグループ毎に単語インデクスを作成する。

検索実行時には、フロントエンドサーバにおいて、検索要求文字列の形態素解析し、単語IDFインデクスから各単語のIDF値を取得し、いずれかの単語を含む全ての検索サーバに対して検索要求を分配する。各検索サーバでは、指定された単語とIDF値に基づいて検索結果をソートし、指定された件数分だけをフロントエンドサーバに返す。フロントエンドサーバは、各検索サーバから返された結果を収集し、マージソートした中から指定された件数分を検索結果としてクライアントに返す。

4.3 インデクス更新

WWW の情報はごく頻繁に更新され、1ヵ月の間に半数以上の文書が更新されるとも言われている。新たに収集した情報を検索サーバに反映する際に、インデクスを改めて作成するか、あるいは既存のインデクスに対して変更分だけを更新するかという二つの方法が考えられる。我々は、頻繁な情報更新を行うことを想定しているため、既存のインデクスを更新することを前提にシステムを構成した。

インデクス更新を行うためには、同一の文書が常に同一の検索サーバに格納されることを保証しなければならない。本システムでは、検索サーバの数に一致する複数のグループに文書集合を分割するが、この時、文書を一意に特定する識別子 (URI やパス名) によって文書をグループ化し、このグループと単語インデクスファイルとを固定的に対応させている。これにより、同一の識別子を持つ文書は、更新後も同一のインデクスファイルに格納されることになる。文書識別子とインデクスファイルとの対応は、参照表を用いて実現している。検索サーバを追加する際には、参照表のエントリを更新し、新たな検索サーバに格納する文書のグループを設定する。

新たなサーバを追加する場合に限らず、参照表のエントリを更新し、対応するインデクスファイルを再構築すれば、ファイルの格納されるインデクスを変更することができる。WWW の中には特に頻繁に情報が更新されるサイトが存在するが、このようなサイトを特定のグループに集約し、このグループについてはインデクスの更新サイクルを短く設定することも可能である。

なお、単語インデクスファイルには、文書の識別子と共に更新時刻を保存しておき、今回収集した文書の更新時刻と比較することによって、変更の有無を判断する。変更が認められない場合は、本文の解析は行わない。

4.4 グローバル IDF の計算

TF×IDF 法を用いた情報検索では、文書集合における単語の出現頻度に基づいてその単語の IDF 値を算出する。本システムの場合、全文書集合における単語の出現頻度に基づいた IDF 値 (以降これをグローバル IDF と呼ぶ) を求めるためには、全ての検索サーバに対応する文書グループから単語の出現頻度情報を収集する必要がある。あるいは、個々の検索サーバ毎に個別の IDF 値を求める方法も考えられる。その場合は一部のグループの更新によって文書全体の IDF 値を再計算する必要もない。しかし、

1. ある検索文字列に対する文書の適合度が、その文書を格納するインデクスファイルとは独立に一定の値を取ることを保証したい
2. インデクスファイルに格納する文書集合を、作為的に一定の条件によって (例えば更新頻度等により) 分割した場合、インデクスファイル間で単語の分布に偏りが生じて検索精度に有意な悪影響を及ぼす可能性がある

という二つの理由により、本システムではグローバル IDF を計算することとした。グローバル IDF 値は、単語 IDF インデクスファイルに格納される。単語 IDF インデクスは単語インデクスと同様に、単語文字列をキーとするトライによって実装されている。単語インデクスでは、単語文字列をキーとして、この単語を含む文書のリストを格納するのに対して、単語 IDF インデクスでは、その単語を含む文書グループと、そのグループでその単語を含む文書数の組のリストを格納する。

いずれかの単語インデクスが更新されると、単語IDFインデクスの該当するグループの値を更新する。更新が完了すると、単語IDFインデクスファイルは情報更新サーバからフロントエンドサーバに複写される。

5 評価

1台のフロントエンドサーバと4台の検索サーバ兼情報更新サーバで構成されるシステムを試作した。いずれもCPU:PentiumII400MHz, RAM:384MB, NIC: 100BASE/TXの同一仕様とした。

このシステム上で、特許公開広報3ヵ月分の文書集合を対象にインデクスを作成し、1単語, 2単語のor, 2単語のand条件のそれぞれについて、文書集合から無作為に単語を抽出し、100通りの検索文字列を生成して、検索速度性能を計測した。表2.に、検索処理全体と、各検索サーバでの検索処理のそれぞれについて、処理時間の平均値を示している。

表 1. 検索対象

検索対象	特許公開広報3ヵ月分
文書数	106252
全単語数	751583
文書サイズ合計	2Gバイト
インデクスサイズ合計	1.4Gバイト

表 2. 検索速度性能 (単位 =msec)

検索パタン	検索時間 (全体)	検索時間 (各サーバ)				ソート時間	その他
1単語	37.8	33.3	31.1	32.5	34.1	0.1	3.7
2単語 or	193.7	189.6	170.1	163.2	179.8	0.1	4.1
2単語 and	250.1	242.5	233.9	246.3	231.3	0.1	3.8

検索時間 (各サーバ) は、4台の検索サーバにおけるそれぞれの実行時間を示している。その他は、(全体の処理時間 − 検索サーバ処理時間の最大値 − ソート時間) により求めた値であり、通信時間を含んでいる。ソート時間は、各検索サーバからの応答を受けた後、フロントエンドサーバ側で行うマージソートに要する時間を示している。全体の検索時間は、最も時間のかった検索サーバの処理時間にほぼ近いものになっており、並列処理の効果が現れていることが確認できる。検索サーバの処理時間以外の要素は検索パタンやヒット件数の値に依らず、ほぼ一定の値となっている。検索サーバの数が n に増加した場合、全体の検索時間は以下のような値に近付くと考えられる。

検索時間 = 検索サーバの処理時間の最大値 + 通信時間 $\times\ n$ + ソート時間 (n, m)

ここで、m はクライアントに提示する検索結果の件数で、ソート時間 (n, m) は、$O(n \log m)$

のオーダとなり、実効的には検索時間に殆ど影響を及ぼさない。したがって、データ規模が拡大し、サーバ数が増大しても検索性能が大きく低下することはないと考えることができる。

6 まとめ

本稿では、WWW の情報空間から収集した文書情報を対象として不特定多数の利用者に対して検索機能を提供する文書検索システムにおいては、スケーラビリティと情報の速報性が重要な要件であることを示し、この要件を満たすために、PC クラスタによる分散サーバ構成を用いた文書検索システムの構築例を紹介した。このシステムは、データ規模の拡張を容易とするデータパーティショニング、頻繁な情報内容の更新を可能とするインデクス更新などの特徴を備えている。小規模なシステムにおける実験結果から、このシステムが実用的な検索性能とスケーラビリティを備えていることを確認した。

参考文献

[1] Web Growth Summary, http://www.mit.edu/people/mkgray/net/web-growth-summary.html

[2] Resource Description Framework (RDF), http://www.w3.org/RDF/

[3] Document Object Model (DOM) Level 1 Specification Version 1.0, http://www.w3.org/TR/REC-DOM-Level-1/

[4] Automatic Text Processing: G. Salton: Addison-Wesley, 1989

[5] Data Structures and Algorithms: A.V.Aho, J.E.Hopcroft, J.D.Ullman: Addison-Wesley, 1983