2U－3

関連文書検索システムの開発(1)　-全体構成-

永峯 猛志† 梅基 宏† 石飛 康浩† 倉持 勉† 倉橋 政之† 増市 博†‡ 舘野 昌一†

†富士ゼロックス(株) IT事業開発部　‡スタンフォード大学 CSLI

1. はじめに

電子化された文書数の増大に伴い、大量の文書からユーザのニーズにあった文書を効率よく検索するシステムが必要になってきている。既存のシステムには全文検索機能[1]を持つシステムがある。全文検索はユーザがキーワードとして指定した任意の単語を含む文書を検索できるが、単語の異表記等の理由により必ずしも満足できる結果を得られるとは限らない。

我々は、このような問題を解決できるシステムとして、大規模分散関連文書検索システムを開発した。本システムは200万件におよぶ日本の公開特許を検索対象としており、ユーザが指定した公開特許をもとに、それに関連した公開特許を検索する関連文書検索機能を備えている。また、本システムで扱う索引は定期的に追加登録が発生するため管理が困難になると考え、索引を複数に分割した。したがって本システムは複数の索引にまたがり検索できる機能を有している。各索引を引くための複数の索引ホストは一台のサーバマシンによって管理されているため、クライアントマシンは、サーバマシンにアクセスするだけで検索を実行できる。以下では、本システムの機能、全体構成および処理時間について述べる。

2. 本システムの機能

本システムの主な機能は次のとおりである。

(a) 関連文書検索

関連文書検索は出現語の統計データに基づいてユーザが指定した文書(入力適合文書と呼ぶ)に関連した文書を検索し、関連度の大きい順に出力する機能である。

(b) 全文検索

全文検索はユーザが指定した単語を含む文書を検索する機能である。

(c) 前方一致検索

前方一致検索はユーザが指定した文字列を先頭に含む単語を持つ文書を検索する機能である。

(d) 複合語検索

複合語検索はユーザが指定した複合語を含む文書を検索する機能である。

関連文書検索についての詳細は"関連文書検索システムの開発(4)—関連文書検索—"で説明する。全文検索、前方一致検索等については、文書構造を考慮した検索が可能である。詳しくは"関連文書検索システムの開発(2)—構造化文書の処理—"で説明する。複合語検索については"関連文書検索システムの開発(3)—複合語辞書—"で説明する。

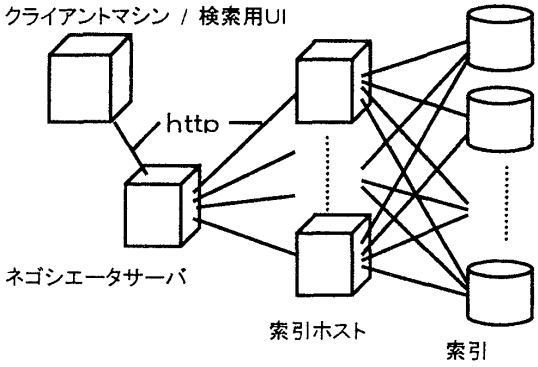

図1:本システムの全体構成

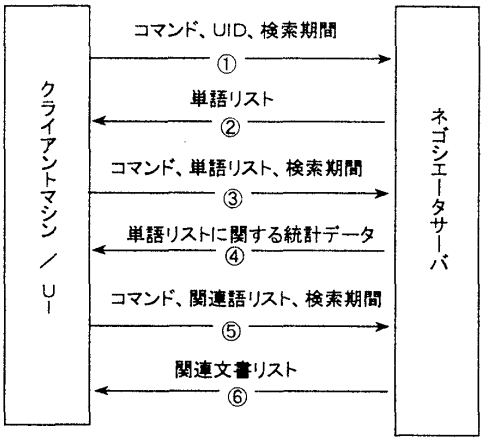

図2:関連文書検索時の動作

3. 本システムの構成

本システムは、UIとしてウェブブラウザを備えたクライアントマシン、クライアントマシンからクエリを受け取り、分散処理を実行するネゴシエータサーバ、実際に索引にアクセスする索引ホストか

Relevant Document Retrieval System (1) -- Overview --

Takeshi NAGAMINE†, Hiroshi UMEMOTO†, Yasuhiro ISHITOBI†, Tsutomu KURAMOCHI†, Masayuki KURAHASHI†, Hiroshi MASUICHI†‡, Masakazu TATENO†

†IT Business Development, Fuji Xerox Co., Ltd., ‡CSLI, Stanford University

らなる(図1)。クライアントマシン、ネゴシエータサーバ、索引ホスト間はhttpを用いて通信する。索引には全文検索用索引、単語検索用索引、複合語検索用索引があり、検索の種類により使い分ける。それぞれの索引は公開特許2ヶ月分(約6万件)で1つの索引を構成している。 次に、図2をもちいて関連文書検索時の本システムの動作を説明する。 図2ではネゴシエータサーバと索引ホスト間の動作については省略してある。

(1) 単語リストの作成: ユーザはUI上で、入力適合文書および検索期間を指定し関連文書検索の開始を指示する。UIは入力適合文書に対応するUID(文書識別子)と検索期間および単語リストを要求するためのコマンドをネゴシエータサーバへ送る(矢印①)。要求する単語リストとは入力適合文書の各文書に含まれる全ての単語である。ネゴシエータサーバは受け取った検索期間をもとに該当する索引ホストへUIDとコマンドを送る。各索引ホストはUIDで示された文書に含まれる単語を単語検索用索引から求め、ネゴシエータサーバへ返す。ネゴシエータサーバは各索引ホストから返された単語リストをマージして、その結果をクライアントマシンへ返す(矢印②)。

(2) 統計データの作成: クライアントマシンは単語リスト、検索期間および単語リストの統計データを計算するためのコマンドをネゴシエータサーバへ送る(矢印③)。ネゴシエータサーバは検索期間をもとに該当する索引ホストを求め、各索引ホストへ単語リスト、コマンドを送る。各索引ホストは単語リストをもとに各単語の統計データを求め、ネゴシエータサーバへ返す。ネゴシエータサーバは各索引ホストから渡された統計データをマージしてクライアントマシンに返す(矢印④)。

(3) 関連文書の検索: クライアントマシンは受け取った統計データをもとに各単語の重要度を計算し、単語とその重要度を対にした関連語リストを作成する。関連語リスト、検索期間、関連文書を検索するためのコマンドをネゴシエータサーバへ送る(矢印⑤)。ネゴシエータサーバは検索期間から該当する索引ホストを求め、各索引ホストへ関連語リストとコマンドを送る。各索引ホストは関連語リストをもとに文書の間連度を求める。索引ホストは求めた関連度とUIDを対にし関連文書リストとしてネゴシエータサーバへ返す。ネゴシエータサーバは各索引ホストから得られた関連文書リストを間連度によってマージソートし、クライアントマシンへ返す(矢印⑥)。クライアントマシンは得られた関連文書リストのタイトルをUI上に表示する。

4. 処理時間

索引ホスト数の変化による関連文書検索の処理時間を図3に示す。 この時間はユーザがUI上で検索の開始を指示し、検索結果である関連文書リストがクライアントマシンに戻ってくるまでの時間である。クライアントマシン、ネゴシエータサーバとしてPC(PentiumII 300MHz、メモリ192Mbyte)を使用した。索引ホストとして同性能のPCを6台使用した。6件の公開特許を入力適合文書とした。

索引ホストの増加に伴う処理時間の増加は目立たない。 これは、並列分散処理が効率よく働いている結果と考えられる。

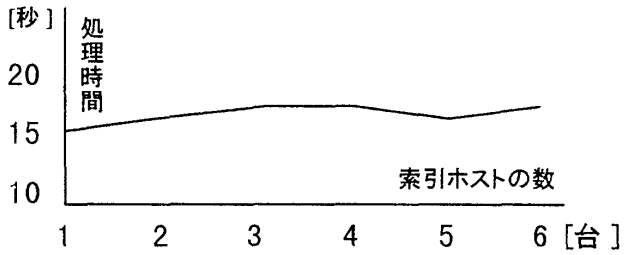

図3:索引ホスト数の変化による検索 の処理時間

5. まとめ

我々は、約200万件に及ぶ公開特許からユーザが望む公開特許を効率よく検索できる大規模分散関連文書検索システムを開発した。関連文書検索機能を用いることにより、従来の全文検索と比較しユーザが望む文書を効率よく得ることができる。本システムは複数の索引を扱うことができるので、検索対象となる文書を追加する場合でも、その文書についての索引を追加するだけでよくメンテナンス性がよい。また、並列分散処理を採用したため複数の索引をもちいて検索した場合でも処理時間の増加が目立たず、実時間で処理できた。

参考文献

[1]増市,山浦,小山,舘野,「形態素解析を用いた全文検索システムとその応用」,情報処理学会 自然言語処理,102-3,PP.17-24(1994.7)