2007 年 3 月 7 日
株式会社日立製作所

約 2 億件の大規模な文書データから瞬時に類似文書を検索できる技術を開発
PCサーバ障害時も継続してサービスの提供が可能なバックアップ機能を実現

日立製作所中央研究所(所長:福永 泰／以下、日立)は、このたび、約 2 億件の大規模文書データから、文書を丸ごと検索キーにして、類似の文書を瞬時に探し出す大規模・高速の文書検索技術を開発しました。日立は 2002 年に、大学共同利用機関法人 情報・システム研究機構 国立情報学研究所(所長:坂内 正夫／以下、国立情報学研究所)と共同で、1000 万件規模のデータベースを対象とした文書検索技術の開発に成功しました*1。今回、既に開発している検索技術と互換性を保ちながら、アーキテクチャを従来の 32 ビットから 64 ビットに拡張した文書検索技術を新たに開発し、検索の対象となる文書量を大規模にした、高速で高精度な検索性能を実現しました。また、大規模データベースに対応するために、複数のサーバを用いて分散処理を行いますが、検索のソフトウェアにサーバの障害を検知する機能をもたせ、障害を生じた場合には直ちにバックアップサーバに切り替えて処理を継続する機能を追加しました。これにより、大規模な文書検索サービスを、24 時間 365 日提供することが可能になります。

情報技術の飛躍的な発展により、あらゆる分野で文書データの蓄積が加速しており、それらの文書データを利用するために、高い精度で文書を検索することが求められています。従来から用いられているキーワードでの検索法は、探したい文書に関するキーワードの見当がつく場合には有効ですが、そうでない場合には、さまざまなキーワードを組み合せて試すなど、所望の文書を得るまでの試行錯誤に多大な労力と時間を要します。これに対し、文書を丸ごと検索キーとして用いる連想検索法では、キー文書に含まれる特徴的な単語を自動的に抽出し、それらの単語の出現頻度をもとに、キー文書との類似度を計算して、大量のデータベースの中から類似文書を検索することができます。

ユーザがキーワードを考える手間がなく、高い検索精度が期待できる文書検索法ですが、大規模データベースで利用するには、計算量が大きいことが課題でした。この課題に対し、日立と国立情報学研究所は共同で、2002 年に 1000 万件規模のデータベース向けに、32 ビットアーキテクチャによる分散処理型の文書検索技術の開発に成功し*2、これまで、学術文献や特許の検索に利用されてきました。しかし、今後、情報量が増えていく中で、文書データベースのさらなる増加に対応すると同時に、所望の文書を瞬時に表示できる検索精度の向上、さらには、システムの耐障害性向上が求められます。

このような背景のもと、日立は、64 ビットアーキテクチャを活用することで、大規模な文書データベースに対応可能な文書検索技術を新たに開発しました。開発した文書検索技術の特長は、以下のとおりです。

1．64 ビットアーキテクチャを用いた大規模･高速･高性能の連想検索技術

アーキテクチャの 64 ビット化により、32 ビットアーキテクチャの限界であった 2 ギガバイトのメモリ空間を超えて、システムのメモリおよびディスク容量の許す限り大規模な文書群を対象に、従来の 32 ビット版と同程度の速さで検索することが可能になりました。また、検索に用いる索引語データベースの機能を強化することにより、より高度な検索が実現可能となりました。

なお、本 64 ビット化は、国立情報学研究所と技術交流をしながら推進しました。

2．連想検索の分散処理に対応した耐障害機能の搭載

大規模データベースに対応するために、複数のサーバを用いた分散処理を行いますが、サーバ数が増えるとハードウェア障害の可能性も高まります。本技術では、検索エンジンのソフトウェアにサーバの通信障害を検知する機能をもたせ、障害を検知した時には直ちにバックアップサーバに切り替えて、検索処理の継続を可能としています。

今回開発した連想検索技術は、社内文書や学術文献、特許をはじめとする大規模な情報データベースを対象とした情報検索のサービスや、放送通信融合時代の番組コンテンツ情報の検索ツールとして、24 時間 365 日のサービス提供を必要とする場面での利用に道を拓くものです。

なお、本成果は、2007 年 3 月 6 日から早稲田大学(東京)で開催される情報処理学会全国大会で発表します。

*1 独立行政法人情報処理振興事業協会(IPA)が実施した「独創的情報技術育成事業」に参画して得られた成果です。

*2 2002 年に日立と国立情報学研究所が開発した連想検索エンジンは、国立情報学研究所のホームページで公開されています。特長は以下のとおりです。

①キー文書に含まれる特徴的な単語を自動選出し、それらの出現頻度に基づく統計量を用いて類似文書を検索。キーワード検索に比べ致命的な検索漏れを防ぐと同時に、内容的に関連する文書に絞り込むことが可能。

②連想検索に用いる索引データを圧縮し、計算機の主記憶上に配置して高速計算。10～20 万件規模のデータベースでも、パソコン上で利用可能。

③1000 万件規模のデータベースに対応するスケーラビリティを得るために、複数のサーバ上に索引データを分散配置して検索を実行。

■照会先

株式会社日立製作所 中央研究所 企画室 [担当:花輪、木下]
〒185-8601 東京都国分寺市東恋ヶ窪一丁目 280 番地
電話 042-327-7777 (直通)

以上