3D-1

連想検索エンジンのスケーラビリティおよび障害耐性の向上

安田 知弘† 今一 修† 岩山 真† 丹羽 芳樹†

(株)日立製作所 中央研究所†

1. はじめに

電子文書はビジネスにおいても教育・研究の場においても不可欠であり、膨大な量の電子文書が日々作成されている。インターネットの拡大も電子文書の量を爆発的に増加させる要因となった。これらの電子文書を最大限に活用するためには、欲しい文書を短時間で検索する文書検索技術が必須である。

最も典型的な文書検索方法は、指定されたキーワードを含む文書を探し出すキーワード検索である。キーワード検索では、所望の検索結果を得るために適切なキーワードを与える必要がある。しかし、未知の文書を検索する際に、その文書に含まれるキーワードを事前に把握できるとは限らない。したがって、適切なキーワードがわからない場合、様々なキーワードを用いて試行錯誤することになる。これは、時間と手間がかかる上に、欲しい文書に到達できない可能性もある。

この問題を解決するために開発されたのが、連想検索技術である[1]。連想検索では検索質問として文書群が入力されると、それらの文書に類似する文書が、あたかも連想されたかのように発見される。本稿では、連想検索技術の概要と、本研究で開発した連想検索エンジンによる連想検索のスケーラビリティ向上および障害耐性機能について述べる。

2. 連想検索技術

連想検索は、ユーザが与えた文書群を検索質問とし、それらに類似する文書を発見する技術である。探したい文書に含まれるキーワードが不明であっても、その文書に関連しそうな既知の文書を検索質問として与えれば、検索が可能となる。

連想検索は、２つのステップからなる。最初のステップでは、検索質問として与えられた文書群に特徴的な単語群を抽出する(図１)。このとき、もとのテキストを直接読むのではなく、各文書に含まれる単語を記録した forward index[2]を使用することで、検索処理を高速化できる。次のステップでは、抽出した単語群に基づき、通常の inverted index[3][4][5]を用いた検索を行なう(図２)。検索結果は、単語頻度等を用いて計算したスコアに従い出力する。連想検索では、特徴的な単語群として多数の単語を抽出するため、特に２ステップ目で計算の負荷が大きくなる。したがって、実用的な速度で連想検索を行なうためには、高速な検索エンジンが必要になる。近年では大容量メモリを搭載したサーバが以前より安価に入手できるため、オンメモリのインデックスを使用して検索を高速化できる。１台のサーバではメモリ容量の制限によりオンメモリインデックスを構築できない場合でも、インデックスを図３のように分割し分散処理を行なえば、エンタープライズサーチ規模のデータであればオンメモリのインデックスが使用できる

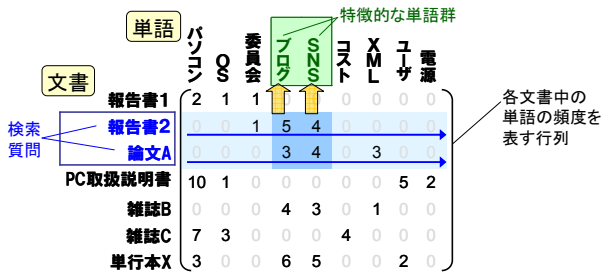

図１ 検索質問文書から特徴的な単語を抽出

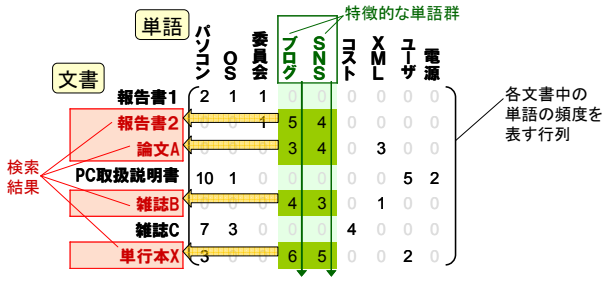

図２ 特徴的な単語が共通する文書を検索

Associative search engine for huge text datasets with fault tolerance

†Tomohiro Yasuda, Osamu Imaichi, Makoto Iwayama, and Yoshiki Niwa

†Central Research Laboratory, Hitachi, Ltd.

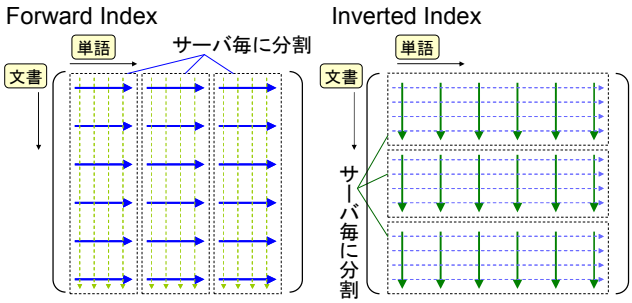

図３ 分散処理のためのインデックス分割

場合が少なくない。高野等が開発した汎用連想計算エンジン GETA (Generic Engine for Transposable Association) 第３版 [1]は、複数のサーバ上に分割して配置したインデックスを用いて、大規模な文書群に対し高速な連想検索を実現している。GETA 第３版は、数千万件・数十 GB オーダの文献集合に対し、連想検索を実行することができる。

3. スケーラビリティの向上

本研究では、電子文書量の更なる増加に対応すべく、GETA 第３版の技術をベースとする連想検索エンジン MANTA (Multi-purpose ANalysis for Transposable Association)を開発した。MANTA は、64bit アーキテクチャでも動作する。単語や文書を識別するために割り当てられる整数値とポインタ長が 64bit 化されたため、事実上、メモリおよびディスクの許す限り大規模な文書群を処理できる。MANTA の開発により、連想検索の対象文書数・単語種類数は、１桁以上向上した。

4. 障害耐性機能

MANTA は高速な検索処理と処理可能な文書量の拡大を目的として、GETA 第３版と同様の分散処理を実装している。しかし、実際の運用では、スケーラビリティの向上に伴いサーバ数が増えるにつれて、ハードウェア障害等により一部のサーバが検索処理を実行できなくなる危険性も増す。このような状況を想定し、MANTA には障害耐性機能が実装されている (図４)。

MANTA で障害耐性機能を使用する際は、通常の検索サーバの他に、バックアップ用サーバを準備する。検索処理時に、あるサーバが応答しない場合は直ちに、そのサーバが行なう予定であった検索処理をバックアップサーバに実行させる。処理させようとしたバックアップサーバも応答しない場合には、他のバックアップサーバを呼び出して処理させる。MANTA はインデックス作成時に、バックアップサーバを通常の検索サーバの代替として動作させるために必要な forward index、inverted index および、検索結果のランキングを行なうために必要な統計情報を、予めバックアップサーバに格納する。バックアップサーバを使用中にもとの通常のサーバが復旧すれば、次回の検索からは、復旧したサーバでの検索を再開する。

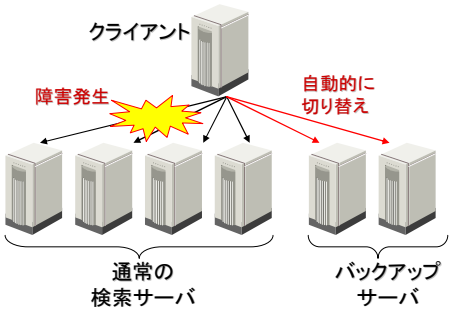

図４ 連想検索エンジン MANTA の障害耐性機能

5. まとめと今後の課題

連想検索エンジン MANTA の開発により、連想検索が可能な文書量は大幅に増加した。また、障害耐性機能の実装により、長期間稼動する検索サービスのエンジンとして用いる場合でも、安定的な運用が容易となった。今後は処理性能やメモリ消費量の定量的評価を行なった上で、さらなるスケーラビリティ向上およびメモリ利用効率向上を図りたい。

6. 参考文献

[1] 高野明彦 他、汎用連想計算エンジンの開発と大規模文書分析への応用, 第 19 回 IPA 技術発表会 (2000).

[2] Page, L. and Brin, S.: The anatomy of a large-scale hypertextual web search engine, Proc. 7th Intl. WWW Conf., pp.107-117 (1998).

[3] Zobel, J. and Moffat, A.: Inverted Files for Text Search Engines, ACM Comput. Surv., Vol.38, No.2 (2006).

[4] Witten, I.H., Moffat, A., and Bell, T.C.: Managing Gigabytes (2nd Ed.): Compressing and Indexing Documents and Images. Morgan Kaufmann, San Francisco (1998).

[5] 北研二，津田和彦，獅々堀正幹 著，情報検索アルゴリズム，共立出版，東京 (2002).