寶珍 輝尚

類似度と索引を工夫して精度良く高速に検索します

時系列データの高速類似検索

■キーワード
データベース　マルチメディア　データベース管理システム　ソフトウェア　ビジュアル言語
データ工学　時系列データ　類似検索　部分一致検索　類似度

■研究の概要

計算機技術の急速な発達に伴い、数値･テキストのみならず画像、音楽や動画といった様々なデータが計算機で扱えるようになってきています。しかし、膨大な量のデータの中から、自分の望むものを的確に、かつ、迅速に見つけ出すことは困難です。このようなものの中に、株価の変動、気温の変化や科学実験の測定データなど時系列の形式をとるものがあります。時系列データの高速な検索に関しましても様々な研究が行われてきています。本研究も、このような時系列データの高速な検索に関する研究です。

■研究･技術のプロセス／研究事例

（1）プロセス

まず、どのような時系列データを検索対象にするのかを明確化します。これは、精度良く検索するために必要な類似度を考えるということでもあります。つまり、どの時系列データとどの時系列データが似ているかを明確にします。これは、類似検索において非常に重要なことです。

次に、高速検索を可能とするためのインデックス（索引）を作成します。インデックスには高速検索が可能となる情報を格納します。この情報は、主に、ふるい落としを行うために使用します。詳細な情報を格納しておけば精度良くふるい落とすことが可能になりますが、データ数が多いとインデックスのサイズが大きくなりすぎて不利になることもありますので注意が必要です。

また、時系列データの類似検索は、大きく、全体一致検索と部分一致検索に分けられます。全体一致検索とは、扱う時系列データの長さが全て同じで、時系列データ全体が検索対象となるものです。部分一致検索とは、時系列データの一部分を検索キーとして、その部分を包含する時系列データを検索するというものです。技術的には、全体一致検索の方が部分一致検索よりも容易に実現することができます。部分一致検索を必要とする場合は、インデックスにもそれなりの工夫をしておく必要があります。

インデックスが作成できましたら検索が可能になります。使い勝手の良い利用者インタフェースを作成すれば上出来です。

（2）事例

我々は、核融合科学研究所と共同研究を行っており、核融合科学実験で得られる様々な時系列データを対象として類似部分検索手法の開発を行ってきております。核融合科学実験では様々なタイプの時系列データが得られます。1回の実験の期間中で緩やかに変化する時系列データもあれば、非常に短い周期で振動する時系列データもあります。また、非常にデータ数が多いのも特徴です。また、部分一致検索も必要とされています。したがいまして、時系列データの類似検索としてはかなりシビアな性能が要求されています。

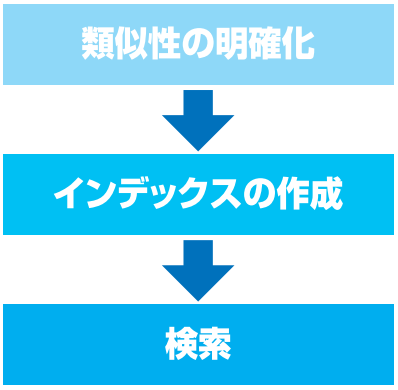

図1　概要

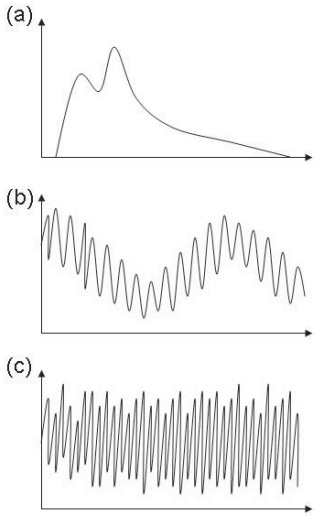

図2　時系列データの例

■研究･技術のポテンシャル

時系列でなくても、数値が列になっているようなデータ（数値列）から類似のデータ（数値列）を高速に求める必要に迫られてはいらっしゃいませんか?同じような手法で解決することが可能です。

■セールスポイント

多くの研究では、類似性の判定にユークリッド距離を使用しています。時系列データによっては、ユークリッド距離の判定では雑音が多すぎて使いものにならないことが多々あります。我々は他の距離を使用することで精度向上を目指しています。