検索語の変動を想定した検索システムに関する研究

山口, 信 / YAMAGUCHI, Makoto

(出版者 / Publisher)
法政大学大学院理工学・工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学・工学研究科編 / 法政大学大学院紀要. 理工学・工学研究科編
(巻 / Volume)
55
(開始ページ / Start Page)
1
(終了ページ / End Page)
2
(発行年 / Year)
2014-03-24
(URL)
https://doi.org/10.15002/00010426

検索語の変動を想定した検索システムに関する研究

STUDIES ON QUERYING TEXT STRING WITH INCREMENTAL DECREMENTAL KEYWORDS

山口　信
Makoto YAMAGUCHI

指導教員 三浦孝夫

法政大学大学院工学研究科電気工学専攻修士課程

In Internet, there are huge amount of information which are mainly text documents such as Blog and Twitter. These information are changes dramatically over time. In this investigation, we propose a new kind of text retrieval,where the pattern changes continuously. Here we introduce effective method of approximate matching for text stream and stream pattern using incremental approach.
Key Words *: Text Stream , Pattern Stream , Approximate Matching*

1.　問題の背景

インターネット上では Blog や Twitter など, 大量の文字列情報が多様な方式で蓄積されている。これらの情報は爆発的に増え続けており、またほとんどの文字列情報は自然言語のまま構造化されていない。このような状況で必要な情報を効率よく検索するためには, 単純な情報検索だけでは対応できない。状況の変化にすぐさま対応し、効率よく検索する手法が求められている。

従来データベース技術によりあらかじめ規定された形式を仮定し, 検索キーを特定の項目に限定することで処理を効率化している。データベースは隣り合うデータの大小関係や、ポインターによりデータ間の関係を表現しているため、その構造は複雑であり、データベースの一部分を変更することは容易ではない。

例えば接尾辞により, 文字列情報を予め解析してデータ構造に展開することで, 検索効率を向上することできる。しかしこれらの構造をオンラインで構築することは非常に高価であり, また基本的に単純一致検索の高速化が中心的課題となっている。

さらに、インターネットやデータストリーム等の, 大容量で変化の激しい環境では, 文字列がオンライン変化しているだけでなく、検索語そのものが移り変わっている。

例えば変動株の最新の変動履歴を検索語として過去の変動履歴を検索すし直ちに類似性を判定することができれば, 変動株に対する素早い対応が可能である。このような状況では、以前に用いた検索語を用いても同様の有意義な検索結果が得られるとは限らない。このため目的の変化に応じて検索語を連続的に変化させる必要が生じる。

このことから情報検索では, 検索語を変動させながら高速に検索を行うアルゴリズムが必要である。本研究では、検索語の解析情報を差分的に作り替えながら行う検索手法を提案する。一度解析した情報の多くは状況が変化してもそのまま使えることが多く、いかにその構造を変化させずに利用するかがポイントとなる。

2.　扱う問題

本研究では, 文字列集合に対する検索の手法を提案する。また最新の情報ほど利用者にとって有意義なものが多いと考え、オンラインでは新しい情報を蓄えるとともに、古い情報は順次削除していくことを想定している。

(1)　変動する文字列

現実世界に存在する情報は時間の変化とともに常に変化を続けている。文字列情報も例外ではなく、インターネット情報技術の発達により大量の文字列情報が爆発的に発生し, 膨大な量の情報から、必要とするデータを検索することが刻々と難しくなっている。データベースに反映されていない最新のデータにこそ重要の情報が含まれているかもしれない。このような状況では、時系列変化した情報をすぐさまデータベースに反映させたり、検索手法の工夫により検索を効率化することする必要がある。

接尾辞は、部分文字列の検索に有用なデータ構造である。中でも接尾辞配列と呼ばれるものは、すべての接尾辞を昇順に並べ、その位置を順に配列に格納したものである。これにより全ての接尾辞の位置を高速に特定することができ、データ圧縮や全文検索など実際に利用される機会が多い。しかし接尾辞の構築は時間空間的に高価であり, 図 1 のように、ストリームデータに対し新たなデータ到着の都度全体の構造を再構築することは容易でない。[1].

一方データベース技術に頼らず、検索手法の工夫により検索効率を改善することができれば、データベースを逐次構築しなおす必要はない。

(2)　変動する検索語

インターネット空間やストリーム等, 大容量で変化の激しい環境では, 文字列の変動だけでなく検索語が連続的に変化する状況が生じる。

利用者の知識が不足していたり、データベースそのものが変化していれば、利用者は目的に対し的確な検索語を入力できるとは限らない。このような状況では、検索結果を反映して新たに検索語を指定することで、より適切な検索を行う

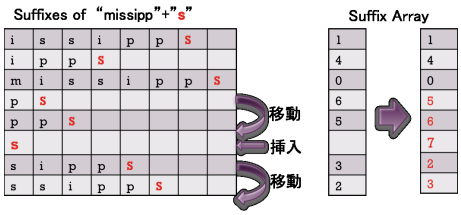

図 1 文字列変化に伴う構造の変化

ことで、検索の精度を高めることができる。図 2 に、検索語を変化させながら行う検索の例を示す。

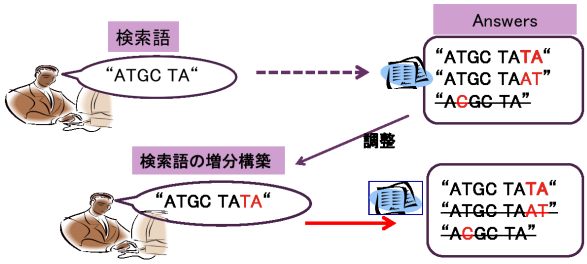

図 2 検索語の変動

KMP 法は検索語構造を工夫することで高速な検索を実現する検索手法である。KMP 法はデータベース技術に頼らず、事前に与えられた検索語を一時的に解析し、冗長な文字列比較を避けながら検索を行う。しかし検索語そのものを変化させる研究は従来なされていない。検索語の変動させながら KMP 法の検索アルゴリズムを適用するためには、検索語表現の差分構築による低コスト化が必要不可欠となる。

本論文では, 検索語の変化に伴い、検索語の解析情報を高速に作り変える手法を提案する。[3].

(3) 近似検索

検索語を目的に沿って変化させる一方で、利用者の意図しない検索語の変化に重要な意味が埋もれている可能性がある。例えば、検索語にタイプミスがあれば必要な検索結果は得られない。他にも図 3 のように、変化し続けるコンピュータウイルスなどは、利用者が感知せずとも検出されるべきである。このような情報を検索するためには、一定の不一致を含んだ近似検索を行う必要がある。

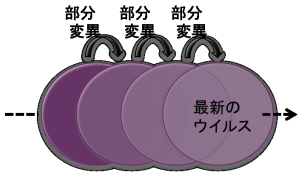

図 3 近似構造を持つウイルス

文字列どうしの一致度合いを表す編集距離は、文字列に対し操作された挿入、削除、置換の回数をエラ-レベルとして定義される。編集距離を用いることにより、エラーレベルが一定値以内のものを検出することができるが、その解析に 2 次元の時空間を必要とする。

本論文では、検索語に変動に伴い、編集距離の解析情報を差分構築する手法を提案する。検索語の変化と近似一致検索により、より柔軟な検索語の解釈を行う。[5].

3. 結び

まず最初に変動する文字列に対する検索の問題について論じた。SuffixAray を動的に構築した結果、1 文字あたり平均 7 ミリ秒分散 8.1 ミリ秒と、どのような文字も均一な実行時間で構築することができた。

次に，検索語を変化させながら行う文字列検索のアルゴリズムについて論じた。KMP 法の解析情報は検索語の増分変化に対し、O(1) 時間で構築できる。ほとんどの場合は新たに解析する必要はなく、解析に要する確率は 10,000 バイト当たり 252 回 (2.5 パーセント) であった。これにより利用者は、目的の変化に伴い検索語の一部を変化させながら高速に検索を行うことができる様になった。

最後に、検索語が変化する状況で行う近似検索について論じた。提案手法では、パターンの増分構築時にポインタなどの予備情報を構築することにより、検索語の減分構築時には 58 倍の性能改善を示した。表 1, 表 2 にロイターコーパスの実験における、差分計算を行わない場合との比較結果を示す。

表 1 増分検索語の実験

文字列長	検索語長	静的構築コスト	動的構築コスト
1	310	604	916
10	310	9914	15994
100	310	93704	161610
1000	310	931604	16684501

表 2 減分検索語の実験

文字列長	検索語長	静的構築コスト	動的構築コスト
1000	310	604	17
990	310	8983	176
900	310	90911	1718
1	310	930673	16499

多くの検索システムは、検索語は利用者が与えるものとして、頻繁に検索語を変える必要性は唱えられてこなかった。クエリ拡張により利用者の意図に沿った検索語を追加するものがあるが、高度に検索語を変化させるものではない。

本手法では検索語の解析情報を差分構築することで検索語の変化にも迅速に対応し、実験により差分構築の速度改善を示した。これにより検索システムは利用者の検索語を柔軟に読み取ることが可能となった。

参考文献

1) 山口信, 三浦孝夫，三浦孝夫: 動的接尾辞配列の実用性, 電子情報通信学会 2012 年総合大会 学生ポスターセッション , 2012.
2) 山口信, 島田諭，三浦孝夫: パターンストリームによるデータ検索, 情報処理学会 第 75 回全国大会 (IPSJ), 2013.
3) 山口信, 島田諭，三浦孝夫: 変動するパターンの文字列検索, 知能ソフトウェア工学研究会 (KBSE) , 2013.
4) Yamaguchi, M. and Miura, T.: Incremental Patterns in Text Search, 7th International Symposium on Intelligent Distributed Computings (IDC2013), 2013.
5) 山口信, 三浦孝夫，三浦孝夫: 変動する検索語の近似文字列検索, 第 6 回データ工学と情報マネジメントに関するフォーラム（DEIM フォーラム) , 2014.