N-gram 全文検索におけるエントロピーを適用した 高次追加インデクス検索語選択方式

非会員 山本 裕* 学生員 佐賀 亮介**
非会員 森山 悟** 正員 辻 洋**

Entropy-Based Incremental Indexing Term Selection Method for N-gram Full Text Search System

Hiroshi Yamamoto*, Non-member, Ryosuke Saga**, Student Member,
Satoru Moriyama**, Non-member, Hiroshi Tsuji**, Member

N-gram indexing is the most popular algorithm for the full text search system where each index consists of serial N characters. Especially the full text search for Japanese text usually has the 2-gram characters index as base. The additional higher-gram index is expected to improve the performance. This paper presents the entropy-based method for mining additional indexing terms from DB in order to reduce the waste of AND operation for 2gram.

キーワード：全文検索, n-gram インデクス, 知識管理, 企業情報システム

Keywords: Full Text Search, n-gram index, Knowledge Management, Enterprise Information Systems

1. 緒 言

日本語全文検索では n-gram インデクス方式, 特に容量の観点から 2gram 方式がよく採用されている⁽¹⁾⁽²⁾が, 2-gram 方式は, 検索語が長い場合にインデクス照合の AND 操作増大により検索効率が悪くなる欠点がある。改善手法として, 3gram 以上の高次インデクスを付加的に作成し検索時に利用する方法があり⁽³⁾, 追加高次インデクス対象語を適切に選択すると検索性能が改善できる。

本研究では, 追加高次インデクス作成対象の検索語を選択する方式として, 文書 DB 中の高頻度語の中から, 2gram の後の文字列出現確率に関するエントロピーを用いた追加高次インデクス対象の検索語選択方式を提案する。つまり, 個別の文書だけでは判らない事項を多数の文書から発見することを旨とするテキストマイニングの研究の一つである。

2. 課題とその解決アプローチ

〈2・1〉 課題 はじめに, 本論で前提とする全文検索方法の概要を図 1 を用いて説明する。本検索方式は, ①ユーザの検索要求後②検索語を高次インデクスファイルと照合し, あれば⑥に, なければ③④検索語を 2gram に分解しインデクスファイルを照合, ⑤各 2gram 照合の AND 操作

を行い, ⑥文書 DB から検索語にヒットする文書を抽出, という流れである。高次インデクスファイルにない検索語に対しては, その照合が余分なステップになるため, 高次インデクスがない場合に比べて検索効率が悪い。

つまり, 高次インデクスは両刃であり, 無秩序に多くの語を高次インデクスとして登録すると性能を劣化させてしまう。語を選択する単純なアイデアはよく検索される語について高次インデクスに登録することであるが, 2gram の AND 操作が負荷にならぬ語を登録してもあまり効果がな

図 1 2gram 及び追加高次インデクス方式概要

Fig. 1. Outline of 2gram and higher-gram indexing.

* 日立製作所 ソフトウェア事業部
〒541-0057 大阪市中央区北久宝寺町三丁目 3-8
Software Division, Hitachi, Ltd.

3-3-8 Kitakyuhoji-cho Chuo, Osaka, 541-0057
** 大阪府立大学 大学院工学研究科
〒599-8531 堺市学園町 1-1
Osaka Prefecture University.
1-1 Gakuen-cho, Sakai, Osaka 599-8531

い。2gram の AND 操作が大きな負荷になる語の選択方法を明確にすることが課題となる。

〈2・2〉 アプローチ AND 操作の負荷が大きくなる語とはどのような語であるかの考察に基づいて課題にアプローチする。はじめに、検索語の文書 DB 中の出現頻度に焦点を絞り、語の頻度 tf 及び同じ語を含む文書頻度 df が多い検索語は、2gram インデックス容量が大きいため AND 操作の負荷が大きいと考える。次に 2gram でヒットするにも拘わらず AND 操作で該当しないと判断されるものが多いとそれは負荷に反映すると考えた。以下では後者の検討を行う。

3. エントロピー値を用いた方式の提案

〈3・1〉 「ムダ」2gram AND 操作 2gram でヒットするが検索語に該当しない AND 操作を「ムダ」2gram AND 操作と定義する。図 2 に検索語と AND 操作 の関係を示す。

検索語 $C_1C_2C_3$ を検索する場合、(1)(2)の 2gram AND 操作が走行するが、(2)は C_1C_2 と C_2C_3 の隣接判定とならず、 $C_1C_2C_3$ に該当せず「ムダ」になる。この場合、 $C_1C_2C_3$ に無関係の②のインデックス部分が大きければ、「ムダ」AND 操作の負荷は大きい。つまり C_1C_2 の後に出現する文字種が多い場合、「ムダ」AND 操作が起こりやすい。従って、「ムダ」AND 操作の負荷が大きい検索語の判断基準は、 C_1C_2 の次に出現する C_k 出現確率の混沌性が高いことである。

〈3・2〉 対象語選択方式 C_k の文字出現の混沌性を評価する指標としてエントロピー値を用いる。連続した文字 C_1C_2 のエントロピー値は次の式で定義できる。

$$E(C_1C_2) = -\sum_i p(C_i|C_1C_2) \log p(C_i|C_1C_2)$$

確率 $p(C_i|C_1C_2)$ は、文字 C_i が文字列 C_1C_2 の次に出現する確率である。 C_1C_2 の次に必ず特定の文字 C_k が出現する場合には $E(C_1C_2)=0$ となり、出現する確率が同じ C_k が複数存在する場合には $E(C_1C_2)=1$ となる。従って、 $0 \leq E(C_1C_2) \leq 1$ であるエントロピー値は、文字 C_k 出現事象の混沌性を示す指標であり、「ムダ」AND 操作の負荷が大きい検索語の選択基準となる。

以上の考察より、追加高次インデックスの対象語つまり「ムダ」AND 操作の負荷が大きい検索語として、(a) tf 及び df が大きい、(b) 2gram に続く文字出現確率のエントロピーが大きい語を選択すればよい。一般にカタカナは、検索語となる頻度が高く、語長が長いことが多い。「イン」「ター」など英語のサフィックス、プレフィックスが多いためエントロピーが大きく、追加インデックス効果が高いと考えられる。

〈3・3〉 具体的検証例 新聞記事 10 万件の中からカタカナで構成され「イン」で始まる検索語のうち、(a) tf が大きい語(b) 各語 3 文字目時点のエントロピーが大きい文字列を対象に提案方式の検証をした。上位 100~1000 語の追加高次(3gram)インデックスを作成・適用し、高次インデックス対象の 3gram を含む検索語での検索パターンを中心に評価し表 1 の結果を得た。表 1 で S1 は「イン」で始まる、tf が大きく 3 文字目時点のエントロピーが大きい 3gram の 1000

図 2 「ムダ」2gram AND 操作と検索語

Fig. 2. Waste of 2gram AND operation.

表 1 追加高次インデックスでの平均検索時間(s/100 件)

Table 1. Average Search Time with 3gram Indexing.

Search Cases	Search Time			
	0	100	500	1000
S1:1000 terms for 3-gram	119	118	88	102
S2:100 terms including 3-gram	124	104	104	101

語、S2 は S1 の 3gram を語の中に含む 1000 語のうち、tf が大きい上位 100 語で構成される検索パターンを示す。

検証例の範囲では、ある程度追加高次インデックスの効果が認められたが、インデックス数、メモリ条件との関係の明確化なども含め、今後本格的な実証評価を行う予定である。

4. 結 言

効果的な追加高次インデックス対象検索語の選択指標として、エントロピー値を用いることを提案した。これは大容量 DB から知識を発見する一つの例と見ることができる。今後は、以下に示すアプローチで本研究を継続する。

(1)エントロピー値を指標とした追加高次インデックスの選択結果の具体的な実験による検証

(2)追加高次インデックスの対象となる検索語を選択する具体的なガイドラインの定義

(平成 16 年 8 月 30 日受付, 平成 17 年 1 月 14 日再受付)

文 献

- (1) K. Matsui, I. Namba, and N. Igata : "Hi-speed Full text Search Engine", *IPSJ SIG Notes Contents Digital Document No.007* (1997)
- (2) Hatakeyama et al.: "ソフトウェアによるテキストサーチマシンの実現", *IPSJ, 情報学基礎研究報告*, Vol.92, No.32, 25-2, pp.19-25,(1992)
- (3) N. Sugaya et al.: A full-text search system for large Japanese text basis using n-gram indexing method, *Proc. 53rd Annual Convention IPS Japan*, 5T-2, 3 (1996)
- (4) Y. Yamamoto, S. Ohmi, and H. Tsuji, , "Entropy-based Indexing Term Selection For N-gram Text Search System", pp.4852-4857, *IEEE/SMC 2003*, October (2003)