5P－11

知識指向文書管理基盤の開発（5）
n-gram 方式に基づく概念検索

松林忠孝† 多田勝己† 菅谷奈津子† 秋沢 充† 後地陽介‡
(株)日立製作所 †システム開発本部 ‡ソフトウェア事業部

1. はじめに

近年、インターネット、イントラネットの利用増大に伴い、生成される電子化文書も爆発的な勢いで増加している。こうした状況下で、大規模な文書データベース(以下、DB と略す)の中から目的の文書を効率的に検索できるシステムに対する要求が益々高まっている。このような要求に応えるために、ORDB 向け全文検索プラグインの拡張検索機能として n-gram 方式に基づく概念検索機能を開発した。本機能によると、探したい情報に関係の深い内容(概念)の文章(以下、種文書と呼ぶ)を入力として検索ができるため、目的の文書を簡単に検索することが可能になる。

本稿では、n-gram 方式に基づく概念検索機能の概要について報告する。

2. 概念検索処理方式の開発

2.1. 開発の目的と課題

従来の概念検索方式では、単語辞書を参照することにより検索に使用する単語(以下、特徴タームと呼ぶ)を種文書から抽出している。しかし、この方式には、次のような問題がある。

(1) 単語辞書の不完全性による検索精度の低下

種文書の中心概念を表す重要語が新語や略語であると、辞書に登録されていない場合がある。このとき、これらの単語は特徴タームとして抽出されないため、目的の概念に沿った検索結果が得られず、検索精度が低下する。

(2) 保守コストが大

辞書に未登録の単語は、これを辞書に登録することで特徴タームとして抽出できるようになる。しかし、そのような特徴タームに対する検索用インデクスの再作成を行わなければ、登録文書中にその特徴タームが含まれている場合でも、検索漏れが発生する。したがって、これを防ぐためには検索用インデクスの再作成が単語登録の都度必要となり、保守コストが増大する。

本概念検索機能の開発においては、上記 2 つの問題を解決することを目的とした。

2.2. n-gram 方式に基づく概念検索機能の特長

今回開発した概念検索機能の特長を述べる。

(1) 重要語の抽出漏れのない高精度な検索

本システムでは、登録文書中から抽出した統計情報を用いて種文書の特徴タームを抽出する。これにより、重要語が辞書に未登録であっても、これを検索に利用した高精度な概念検索を実現可能である。

(2) 優れた保守性

検索アルゴリズムには特徴タームによらず検索漏れの発生しない n-gram インデクス[1]を採用している。このため、検索用インデクスの再作成が不要であり、保守性に優れた概念検索システムを実現可能である。

3. 概念検索処理の概要

本システムの登録処理と検索処理の概要を図 1 に示す。

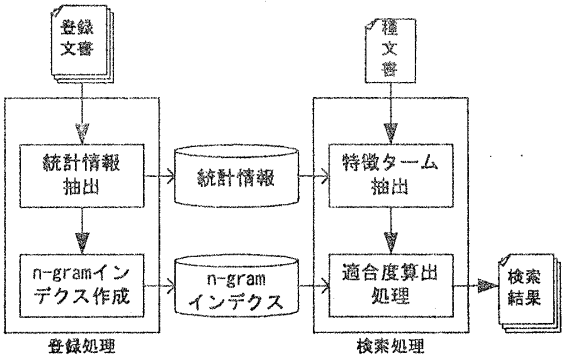

図1 概念検索処理の概要

3.1. 登録処理

登録処理においては、登録文書から所定長の n-

Development of Conceptual Search Method based on n-gram Index.
Tadataka MATSUBAYASHI†, Katsumi TADA†, Natsuko SUGAYA†, Mitsuru AKIZAWA†, Yousuke USHIROJI‡
†Business & Information Systems Development Division, Hitachi, Ltd. ‡Software Division, Hitachi, Ltd.

gram を抽出する。そして、各 n-gram について文字種境界情報や出現文書数情報等の統計情報を抽出し、蓄積する。また、各 n-gram の出現位置情報を n-gram インデクスに登録する。

3.2. 検索処理

検索処理においては、登録処理で蓄積した n-gram の統計情報を用いて種文書から特徴タームを抽出する。これにより、DB に登録されている文書の内容に応じた特徴タームが、単語辞書を用いることなく抽出可能となる。次に、ここで得た特徴タームの出現情報を基に DB に登録されている文書との適合度を算出する。

3.2.1. 特徴ターム抽出方式

“水不足”という複合語が種文書中に存在する場合、その複合語を構成する単語を含む“水が不足している”という文書も概念検索結果として抽出する必要がある。そのため、本特徴ターム抽出処理では、登録文書から抽出した文字種境界情報を用いて、種文書中に存在する複合語をそれを構成する最小単位の単語へと分割する。これは、「単語は助詞や接続詞を伴って使用される可能性が高い」という仮定に基づくものである。本処理の概要を図 2に示す。

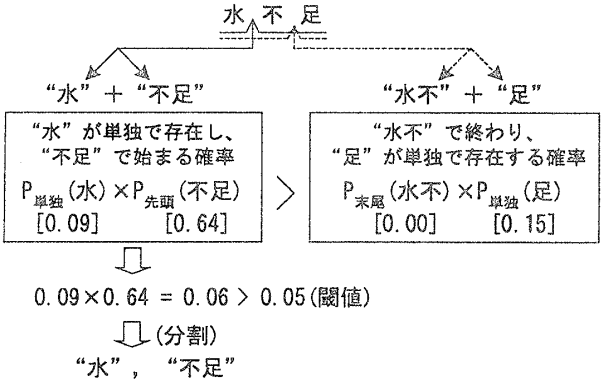

図2 特徴ターム抽出処理の概要

本方式では、先頭から隣り合う２点の分割候補点に対し、文字種境界確率を用いて分割確率を算出する。そして、単語間にまたがる部分文字列が単語として抽出される分割誤りを排除するために、上記２点間の分割確率を比較し、値が高い方の点を選択する。そして、分割確率が所定の閾値より低い場合は単語の区切れ目になる可能性はないものと判断して、そこでの分割は行わない。分割確率が所定の閾値より高い場合は、そこで文字列を分割する。これらの処理を文字列の先頭から末尾まで繰り返すことにより複合語から単語を抽出する。図 2に示した“水不足”の場合、“水不”および“不足”の分割確率は前者の方が大きく、かつ所定の閾値よりも大きいので、この点で分割する。すなわち、複合語“水不足”から“水”と“不足”が抽出される。

3.2.2. 適合度算出処理

適合度算出処理では、各特徴タームに対して TF･IDF (Term Frequency ・ Inverted Document Frequency)モデル[2]に基づいた重み付けを行う。そして、各特徴タームの持つ重みをより多く含む文書に対し高い適合度を付与する。これにより、種文書中に記載されている概念に対する適合度を算出する。

4. 概念検索機能

今回開発した全文検索プラグイン概念検索機能では、基本機能に加え、以下に示す検索機能を実現した。

(1) 構造指定検索

検索の対象とする登録文書の論理構造を指定した概念検索を行う。これにより、文書の論理構造の表す意味を意識した高精度な概念検索を行うことができる。

(2) 同義語・異表記展開検索

辞書ベースの同義語展開処理およびカタカナやアルファベットに対するルールベースの異表記展開処理を行う。これにより、表記の揺らぎを吸収した漏れのない概念検索を行うことができる。

5. まとめ

ORDB 向け全文検索プラグインの拡張検索機能として n-gram 方式に基づく概念検索機能を開発した。これにより、以下の特長を持つ概念検索機能を実現することが可能になった。

(1) 重要語の抽出漏れのない高精度な検索

(2) 優れた保守性

今後、文書管理ミドルウェア DocumentBroker との連携方式を検討していく。

参考文献

[1] 川下他：「構造化文書対応全文検索システム Bibliotheca2 TextSearch の開発(1)～(4)」、情報処理学会第 55 回全国大会 4N-3～6

[2] William B.Frakes 他 ： 「Information Retrieval」 pp363～392, Prentice Hall