

HBase マッピングによる日本語 WordNet を利用した 番組等検索システムに関する研究

A Study on Audiovisual Content Retrieval System using WordNet and HBase Mapping

陳 豊†
Feng Chen

スィープラサスック パオ†
Pao Sriprasertsuk

亀山 渉†
Wataru Kameyama

1. はじめに

地方振興の一環として、その地方に関連した人気キーワードをブログや SNS などから抽出し、それらに合致する番組、スライドショー、及び種々の情報（以下、コンテンツと呼ぶ）の配信をエリアワンセグや Wi-Fi で自動的に行えるシステムを実現するため、本研究では、日本語 WordNet を使用し、人気キーワードをその上位概念と下位概念を使って拡張し、コンテンツメタデータと照合する手法を検討している。本報告では、検索効率を向上させるため、日本語 WordNet を元の形式である SQLite データベースから HBase へマッピングし、上位概念と下位概念のデータを再構築する手法の検討を行った。

2. 提案手法

2.1 提案システム全体像

提案するシステムでは、最初に、日本語 WordNet を使用して、コンテンツメタデータが所属する概念及び上位概念を抽出し、それをデータベースに蓄積する。次に、人気キーワードが所属する概念及び上位概念を抽出し、保存されたコンテンツメタデータの所属概念及び上位概念と照合する。両者の所属する概念が一致しなくとも、同じ上位概念を共有している場合は、人気キーワードとコンテンツメタデータは近い概念関係にあると考えられる。そこで、その人気キーワードとコンテンツメタデータの類似度を検出するため、Wu-Palmer アルゴリズム[1]で類似度計算を行う。得られた結果が指定した Threshold より大きい場合、そのコンテンツは人気キーワードに合致したとし、検索結果として出力する。全体のプロセスを図1に示す。

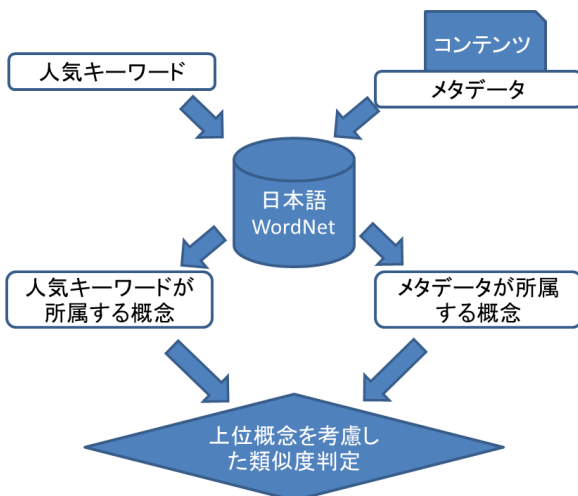


図1 提案システム全体プロセス

† 早稲田大学 大学院国際情報通信研究科

2.2 HBase テーブル設計

本研究ではシステムのスケーラビリティと高速アクセスを実現するため、大規模データに特化した Hadoop と HBase を使用する。公開されている日本語 WordNet はリレーショナルデータベース SQLite に保存されているため、HBase のスキーマにマッピングする必要がある。マッピングのための個々のテーブルのスキーマを表1に示す。

表1 HBase テーブル設計

| Table | Row Key | Column Family | Value |
|----------|-----------|--------------------------------|------------------------|
| lookup | word_id | lang : word_id | word |
| word_syn | word_id | forward-distance : synset_id | synset_exp |
| syn_word | synset_id | lang : word_id | backward-distance word |
| syn_con | synset_id | backward-distance : content_id | url, etc. |

lookup テーブルには個々の単語と id を保存する。日本語と英語を区別するため、Column Family には言語フラグ lang を立てる。word_syn テーブルには単語と概念の関係を保存する。効率的な検索を実現するため、Column Family には、単語が直接所属する概念を forward-distance を 1 とし、また、単語が所属する一階層上の上位概念を forward-distance を 2 とし、該当する synset_id と共に保存する。synset_exp はその概念の内容である。syn_word テーブルには概念と単語の関係を保存する。lookup テーブルと同様に、Column Family には言語フラグ lang を立てる。value には synset_id と word_id の関係を、単語が直接 synset_id に所属する場合は backward-distance を 1 とし、単語が一階層下の下位概念に所属する場合には backward-distance を 2 とし保存する。syn_con テーブルには概念とコンテンツメタデータの関係を保存する。コンテンツメタデータが直接所属する概念と下位概念を区別するため、backward-distance フラグを立てる。

人気キーワードを処理する際には、まず lookup テーブルから対応する word_id を検索し、この word_id で word_syn テーブルから上位概念を検索する。その上位概念によって、関連するコンテンツメタデータ (content_id) を検索する。最後は類似度を計算した後、共通の上位概念を使用し、syn_word テーブルでメタデータを拡張する。

3. 実験

提案したシステムを評価するため、現在公開されている SQLite による日本語 WordNet と比較して、検索機能及び性能を測定した。

3.1 実験環境

今回の実験では、正確な比較のために、Linux サーバ 1 台で提案システムと SQLite データベースを評価する。実験環境を表 2 に示す。

表 2 実験環境

| | |
|--------|--|
| CPU | Intel(R) Core i7 3.33GHz |
| Memory | 12GB |
| OS | CentOS 6.3 |
| Java | Java 1.7 |
| HBase | Hadoop 2.0-CDH 4.1.2, Hbase 0.92-CDH 4.1.2, Pseudo-Distributed |
| SQLite | Package: org.sqlite.JDBC |

3.2 単語から上位概念の検索

日本語 WordNet から単語をランダムに 10 個及び 100 個を選択してキーワードとする。これを検索データセットとして使用し、各キーワードの上位概念を検索する。提案する HBase による検索と SQLite による検索を 5 回行い、平均値を計算する。比較した検索時間を図 2 に示す。

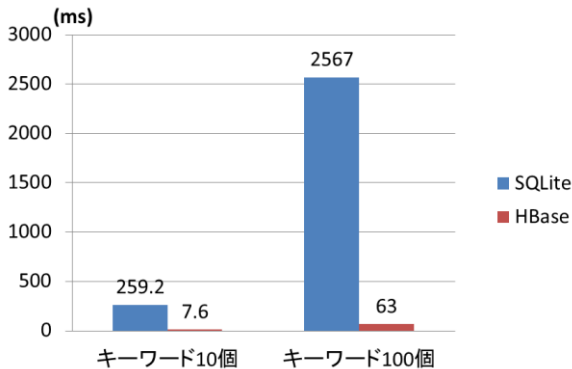


図 2 単語による上位概念の検索時間の比較

10 個のキーワードを検索する場合、提案した HBase による検索時間は約 1/34 に短縮された。また、100 個のキーワードを検索する場合、約 1/40 に短縮された。

3.3 概念から下位概念の単語の検索

日本語 WordNet から概念をランダムで 10 個及び 100 個を選択し、検索データセットとして使用する。これにより、各概念の下位概念に所属する単語を検索する。比較した検索時間を図 3 に示す。

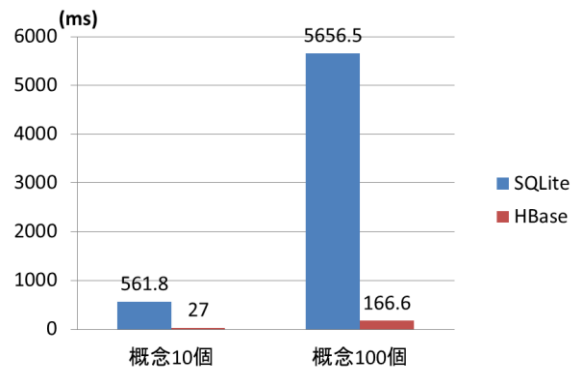


図 3 概念による下位概念の単語検索の比較

10 個の概念を検索する場合、提案した HBase による検索時間は約 1/20 に短縮された。100 個の概念を検索する場合、約 1/34 に短縮された。

3.4 概念から下位概念の日本語の単語の検索

提案したシステムは、検索効率を保持しつつ、下記の検索パターンにも対応できる。

- 1 単語から所属する概念と上位概念を検索する。
- 2 概念から言語ごとに所属する単語と下位概念の単語を検索する。

例として、概念から下位概念の日本語の単語の検索時間を測定した結果を図 4 に示す。

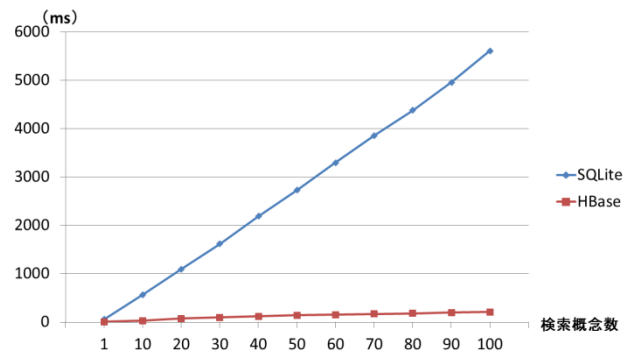


図 4 概念から下位概念の日本語の単語の検索

1 つの概念で検索する場合、SQLite では 56 msec、HBase では 3 msec である。10 個の概念から検索する場合、SQLite は 565.8 msec で、HBase は 47.4 msec である。つまり、提案した HBase の検索時間は約 1/12 に短縮されることが分かった。また、100 個の概念で検索する場合、SQLite は 5609.8 msec で、HBase は 204.4 msec であるため、約 1/27 に短縮される。図 4 から、HBase の検索時間は検索概念数が増加しても、増加の幅は非常に小さいことが分かる。

4. まとめと今後の課題

本報告では、HBase を利用した日本語 WordNet の効率的な検索システムを提案した。

今後の課題としては、提案したシステムの構築を完了し、実証実験を行うと共に、評価を行う予定である。

謝辞

本研究の一部は、総務省「戦略的情報通信研究開発推進制度 (SCOPE)」(採用課題番号: 122304003) の研究助成によるものである。ここに記して謝意を表す。

参考文献

- [1] Wu, Z., Palmer, M., "Verb Semantics and Lexical Selection", 32nd Annual Meeting of the Association for Computational Linguistics, pp. 133-138, New Mexico State University, Las Cruces, New Mexico (1994)