4U-06 マーケットバスケットデータに対する類似索引機構の文献データへの適用*

三浦 貴之 大保 信夫 古瀬 一隆†
筑波大学大学院 工学科

1 はじめに

文献検索は様々な分野で必要とされており、多くのシステムにおいてキーワード検索システムを用いられている。個々の文献はその内容を代表するキーワードが付けられており、キーワード検索システムでは、付加されているキーワードをクエリ中で指定して検索をする。しかし、文献に付加されているキーワードを特徴とした特徴ベクトルにおいて、文献の類似検索を行うと、次元数が増加するにつれて検索効率が低下するという問題が起こる [2]。

文献データと同様に多くの特徴を持つデータであるマーケットバスケットデータに対し、Charu C. Aggarwal らがシグネチャテーブルと呼ばれる索引を用いた類索引検索の手法を提案している [1]。

本研究は、マーケットバスケットデータと文献データの特徴に着目し、シグネチャテーブル法を用いて、文献データの類似検索の効率向上を試みる。

2 マーケットバスケットデータに対する類似索引機構と文献データへの適用

[1] で Charu C.Aggarwal らの提案した手法 (以下 シグネチャテーブル法と呼ぶ) は、マーケットバスケットデータの類似検索を効率的に行うために必要とされる、インデックスの手法の一つである。

シグネチャテーブル法は、多次元特徴ベクトルで表現されているマーケットバスケットデータをビット列 (K ビット) で表すことにより、全トランザクションを 2^k のエントリを持つ索引に振り分ける。類似検索を行うに当たり、索引付けされたトランザクションに対しクエリを発行し、Branch and Bound Technique を用いて絞り込みを行う。

文献データへの適用と問題点

キーワードを文献の特徴とし、マーケットバスケットデータと同様に、文献データを多次元特徴ベクトルとし検索を行う。

シグネチャテーブル法を用いて文献データの類似検索を行った。クエリをデータベーストランザクションの中から選び、 1000 回発行し、k-NN を探し出すまでに全体のデータベーストランザクションを調べる割合を計算して平均をとる。文献データとして JICST95(全文献数 40000, 全キーワード数 16733) を用いた。また類似関数はマッチ (文献が共通に持つキーワードの数) を用いた。

10-NN サーチを行ったところ、図 2(ALL KEY WORD) に示すようにデータベーストランザクションの約 80%を調べなければならない。

問題点

マーケットバスケットデータにおいての、トランザクションとアイテムの割合に対し、文献データでの文献数に対するキーワードの数が多いという特徴がある (表 1)。このようなデータに対し、シグネチャテーブル法を適用すると次のような問題が起こり、検索効率を悪化させる原因になる。

1 つのシグネチャ S_i に 多くのキーワードが入っていると、クエリが S_i と共通に持つキーワードと、データベーストランザクションが S_i と共通に持つキーワードが一致しにくくなり、“フォルスドロップ” が起こりやすくなる。フォルスドロップが起こりやすくなることにより、pessimistic bound に対し optimistic bound が大きくなり、検索効率が悪化する。

この問題を解決するために全キーワード集合を分割する数を多くすると、エントリの数が膨大になり、各エントリとの optimistic bound の計算やソートに時間がかかる。そこで本研究ではキーワードを減らすことにより、一つのシグネチャに入るキーワード数を減らす。

	MB データ	文献データ
アイテム (キーワード) 数	1,000	16,733
トランザクション (文献) 数	40,000	40,000

表 1 文献データとマーケットバスケットデータの比較

3 キーワード削除法

キーワードを減らす際に、どのようなキーワードを削除するのかが問題である。本研究では、キーワード間の関連度に着目し、「他のどのキーワードにも関連度が

[1]Exploiting the Branch and Bound Technique for Document Similarity Search
[2]Takayuki Miura, Nobuo Ohbo, Kazutaka Furuse

低いキーワード」を削除の対象とし、残ったキーワードを基に類似検索を行う。その為にキーワード間の関連度である “confidence” を定義する。

confidence とキーワードの削除

まず、キーワード間の関連度として “confidence” を以下のように定義する。

$$conf(x,y) = \frac{x \text{ と } y \text{ が同時に出現する文献数}}{x \text{ が出現する文献数}} \times 100$$

次に、キーワードの削除法について以下に述べる。1つのキーワードに対して1つのノードに割り当て、各ノード間には両端のノードの confidence を重みとした枝を持つ有効グラフで表し、キーワード x のノードが以下の条件を満たすときキーワード x を削除する。
閾値 R(%)、任意のキーワード y ($\neq x$) に対し、

$$(conf(x,y) \leq R) \vee (conf(y,x) \leq R)$$

図1は閾値を 40%にしたときのキーワードが削除される様子である。

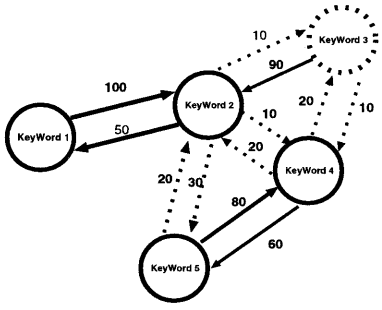

図 1: **キーワード削除法 (閾値 R=40(%))**

実験

文献データ JICST95 を用いて、キーワードを減らし、10-NN サーチを行った。図2では閾値を 10%、15%、18%、と変化させて 10-NN サーチを行った結果である。
閾値を上げるほど検索効率が良くなっていることが分かる。confidence に基づきキーワードを減らすことにより、フォルスドロップが起こりにくくなっている。それによって、特定のエントリとの optimistic bound だけが高く、その他のエントリとの optimistic bound は低くなり、絞り込みの効率が良くなった。
キーワードを減らすことによって類似検索の効率を向上させることが出来た。しかし、文献間の類似度を共通に持つキーワードの数としているので、検出される結果の文献が キーワードを減らす前の結果と異なることがある。そこで 10-NN サーチを行い、キーワードを減らす前と後での検索結果がどのぐらい一致しているかを図3に示す。
confidence 10%においてキーワードを削除し類似検索を行ったところ、その結果はキーワードを減らす前の検索結果と 90%以上一致する結果となり、検索結果において信頼性の高い結果を得ることが認められている (図3)。

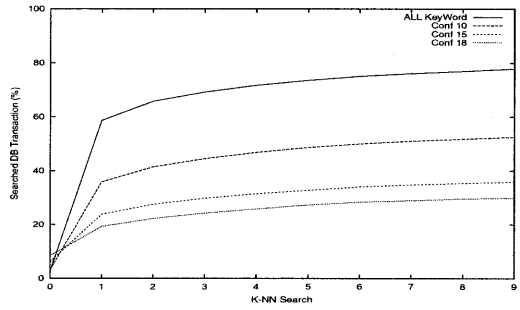

図 2: **K-NN サーチ**

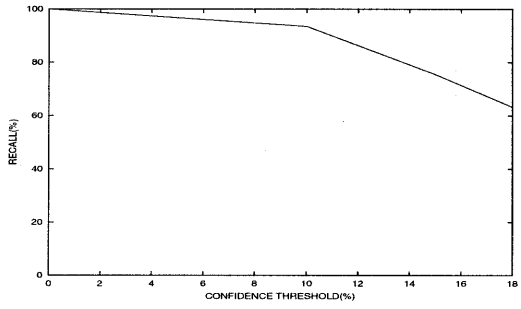

図 3: **Recall - Confidence Threshold(%)**

4 まとめ

本研究では、マーケットバスケットデータに対する類似索引機構を文献データに適用し、文献データのキーワード数を効果的に減らすことにより、類似検索の効率向上が見られた。今後の課題として、confidence の閾値の決め方や、検索結果の信頼性を検討する必要がある。また、他の文献類似度を用いての類似検索や、類似度に反映したキーワードの削除法も検討中である。

参考文献

[1] Charu C. Aggarwal, Joel L. Wolf, Philip S. Yu . A New Method for Similarity Indexing of Market Basket Data.SIGMOD '99 Philadelphia PA,1999

[2] Roger Weber, Hans-J. Schek, Stephen Blott: A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. Proceeding of the 24th VLDB Conference New York, USA, 1998