6E－9

シグネチャファイルによるスペクトルデータベース特徴検索システムの構築

包 赤軍† 北川 博之†† 鈴木 功††

†筑波大学 理工学研究科 ††筑波大学 電子・情報工学系

1 はじめに

化合物のスペクトルデータ検索は未知化合物の同定等を始めとして多くの化学応用において必要とされる[1]。スペクトルデータ検索では、通常のデータベース検索といくつかの点で異なり、検索用入力データとデータベース中のデータが完全に一致する必要はなく、通常はユーザが入力した複数の既知ピーク波数、強度、許容誤差から、類似度の高い順に指定されたx個のスペクトルを検索することが要求される。またピーク波数の指定では、吸収帯のない領域を指定することもしばしばある。

我々は、従来テキスト検索で用いられてきたシグネチャファイルをスペクトル特徴検索に応用するために、二次元のシグネチャファイルの概念を提案した[2]。本稿では、実際の赤外線スペクトルデータ集の一つであるIRDCデータベースを対象として、スペクトルシグネチャファイルを作成し、その有効性の評価を行なったので報告する。

2 スペクトルシグネチャファイル

シグネチャとは、個々のデータオブジェクトから生成される固定長のビット列である。このシグネチャとデータオブジェクトの識別子の組を格納したのが、シグネチャファイルである。また、シグネチャファイルの構成法はいくつかあるが、本研究では、ビットスライスシグネチャファイル(Bit-Sliced Signature File, BSSF)を用いた。BSSFでは、シグネチャがビットごとに別々のファイルに格納されるため、BSSFのうちの一部にアクセスすれば良いので、一般的に検索コストは小さくて済む。スペクトルシグネチャファイルを作成する手順は次の通りである。

1. 図1のように、一定の波数幅wで、波数領域を分割する。1小領域はシグネチャの1ビットと対応させ、小領域内ピークの有無により、対応するビットには、“1”または“0”をセットし、要素位置シグネチャを生成する。また同じように強度値から要素強度シグネチャを生成する。
2. 各要素シグネチャを1ページに格納されるスペクトル毎にグループ化し、スーパーインポーズドコーディングにより、ページシグネチャを生成する。
3. BSSFを用いてページシグネチャとページ識別子を格納する。

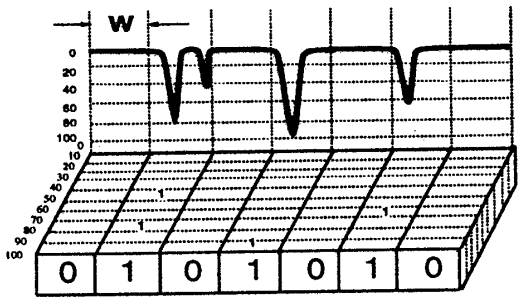

図1. スペクトルに対応する要素シグネチャの生成

3 検索アルゴリズム

Step 1: ユーザが許容誤差、ピークの位置と強度データ、検索結果として出力すべきスペクトル数xを入力する。それに基づき問い合わせシグネチャを作成する。

Step 2: 問い合わせシグネチャとBSSF中のページシグネチャの位置シグネチャのマッチングを行う。マッチングが成功する全てのページについて、次の処理を繰り返す：

(1) 以下の条件のいずれかが成立するか調べる。

1. 既に選択されているスペクトル数がx個未満である。
2. 既にx個のスペクトルが選択されており、このページの強度シグネチャから算出された類似度がx個のスペクトルの類似度の最小値Minより大きい。

(2) いずれかの条件が成立する場合、このページ中に含まれるスペクトルデータを調べ、実際には検索条件を満たさない、フォルスドロップと呼ばれるスペクトルデータを取り除く。そして検索条件を満たすスペクトルデータを、これまでに選択したものと合わせて、類似度の大きい順に、最大x個のスペクトルを選択し、またその類似度の最小値をMinに置く。

Step 3: 選択されたスペクトルのID番号、化合物名、分子式とそれぞれの類似度を出力し終了する。

4 シグネチャファイルの性能評価

フォルスドロップの生じる確率はシグネチャファイルの性能を評価する上で重要な尺度である[3]。シグネ

Development of a Spectra Retrieval System Using Signature Files

C. Bao, Master's Degree Program in Science and Engineering, Univ. of Tsukuba

H.Kitagawa and I.Suzuki, Institute of Information Sciences and Electronics, Univ. of Tsukuba

チャの長さが長い方が一般的にフォルスドロップ確率は小さくなるが、シグネチャファイルの格納コストが大きくなってしまう。従って、フォルスドロップ確率が一定のレベルで、格納コストをできるだけ小さくするのが望ましい。本稿では、フォルスドロップ確率を次の三種類に分けて議論する。各変数は表1のようになっている。

1. 位置シグネチャのみを用いて検索する場合のフォルスドロップ確率 F_{d1}
$$F_{d1} = \frac{P1-A1}{D-A1}$$
2. 強度シグネチャを用いる上でのフォルスドロップ確率 F_{d2}
$$F_{d2} = \frac{P2-A2}{D2-A2}$$
3. 検索システムの総合フォルスドロップ確率 F_d
$$F_d = \frac{P-A}{D-A}$$

Fd	フォルスドロップ確率
A	出力するX個のスペクトルを含んでいたページ数
A1	ピーク位置条件を満たすスペクトルを含んでいたページ数
A2	P2の中でそのページを調べた時点では、問い合わせ結果に含めるべきスペクトルを含んでいたページ数
P	実際にアクセスしたページ数
P1	位置シグネチャがマッチング条件を満たしたページ数
P2	D2の中で強度シグネチャがマッチング条件中で満たしたページ数
D	スペクトルデータを格納するために必要なページ数
D2	X個のスペクトルが選択された時点以降に、位置シグネチャがマッチング条件を満たしたページ数

表1. 変数のリスト

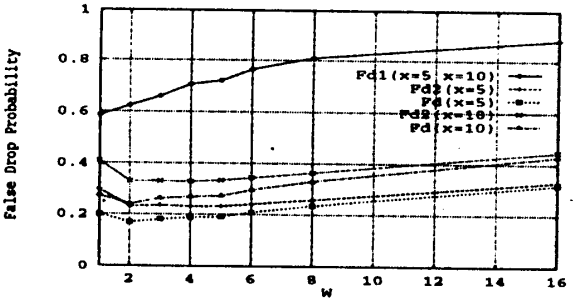

図2. フォルスドロップ確率
(シグネチャファイルサイズがソースデータの30 %)

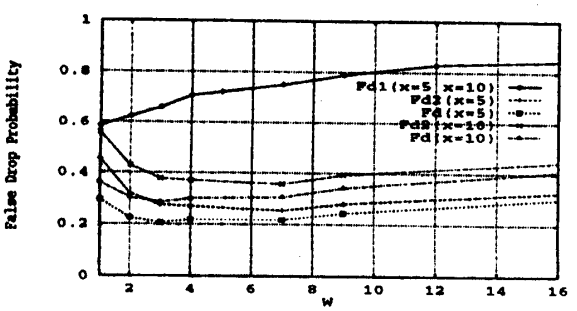

図3. フォルスドロップ確率
(シグネチャファイルサイズがソースデータの15 %)

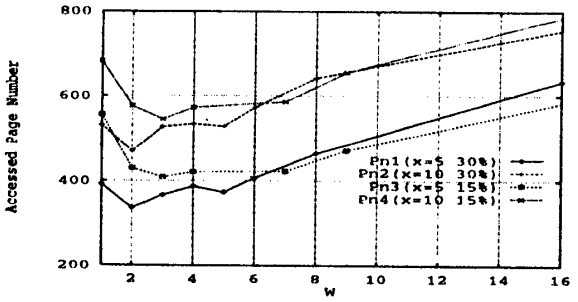

図4. アクセスするページ数

シグネチャファイルのサイズは波数幅、強度幅と1ページに格納するスペクトルの数により決まる。一つのスペクトルデータ当り約400Bの記憶容量が必要である。1ページを4KBとすると、10件スペクトルデータを格納することができる。本研究では、シグネチャファイルサイズをソースデータの15 %及び30 %とした場合について、波数幅wとフォルスドロップ確率の関係(図2、3)及び検索する時にアクセスされたページ数との関係(図4)を測った。波数幅wが大きくなると F_{d1} は単純増加する。一方、wが大きくなるに伴い強度幅が小さくなるため、最初 F_{d2} が減少するが、F_{d1} の増加に伴い F_{d2} も増大する。各シグネチャファイルサイズに対して、wが2 ~ $4cm^{-1}$ の範囲に、F_d とアクセスしたページ数が最小になる点がある。シグネチャファイルの大きさがソースデータファイルサイズの30 %の場合、w=$2cm^{-1}$ において、F_d が0.17以下になる。また、xが増加するのに伴い F_{d2} も増加し、結局は F_d とアクセスしたページ数も増加する。

5 まとめと今後の課題

本稿では、スペクトルデータベース特徴検索を対象としたシグネチャファイルの構成法について検討した。その結果、強度シグネチャを用いることの有用性を確認した。今後の課題としては、吸収帯のない領域を指定した検索条件への対応とスペクトルデータをページに格納する場合のクラスタリングの方法の検討などがあげられる。またピークの不均一な分布を考慮しシグネチャ生成時の波数幅を調整する方法を検討することも必要である。

参考文献

[1] 田辺 和俊, 田村 禎夫, 佐伯慎之助, 鈴木 功, 田隅 三生, "赤外スペクトルファイル検索システム", 分光研究 第32巻 第4号 (1983).

[2] 北川 博之, 包 赤軍, 鈴木 功, "シグネチャファイルに基づくスペクトルデータベースの検索方式", 電子情報通信学会1995年総合大会 (1995).

[3] H.Kitagawa, Y.Fukushima, Y.Ishikawa and N. Ohbo: "Estimation of false drops in set-valued object retrieval with signature files", Proceedings of the 4th Intl. Conf. on Foundations of Data Organization and Algorithms (FODO), Springer-Verlag, LNCS 730, pp.146-163 (1993).