

機械可読化のためのポリマーデータの単位統一と数値変換

Unit Unification and Value Conversion for Machine-Readable Dataset Creation

物材機構 ○岡 博之, 石井 真史

NIMS ○Hiroyuki Oka, Masashi Ishii

E-mail: OKA.Hiroyuki@nims.go.jp

1. 序

NIMS ではポリマーデータを学術論文から収集しており、それらをデータベースで管理・公開し、また、マテリアルズ・インフォマティクス(MI)に活用している^[1]。データ収集は人手で行っているが、最近論文出版数の急増による人手作業増やコンピュータによる文書処理技術の向上などを考慮して、コンピュータを用いた自動抽出も検討しており、その結果、論文中の表からポリマーデータを大量に抽出できるようになった^[2]。この自動抽出では物質名(ポリマー名)・物性名・物性値の3つ組を1データとして抽出しているが、物性値では単位がミリやマイクロ付きの数値となっていることが多く、データを MI で利用するときは統一した単位への数値変換が必要となる。また、同オーダーの単位でも、文中の表現が”g/cm³”や”g cm⁻³”のように文字列的に多様であることも多い。そこで本研究では、抽出したデータの単位を認識し、そして、統一した単位へ数値変換することを検討した。抽出した物性値には”0.21±0.01”や”1.1×10⁻⁵”のように誤差や累乗数が記載されている場合があり、このようなときは事前に数値のクリーニングを行う必要がある。本研究ではこの処理についての結果も報告する。

2. 実験

2-1. 使用データおよびプログラム言語

データは、表から自動抽出したポリマーデータを用いた。その技術詳細は以前に報告の通りである^[2]。ポリマー物性は主要なものだけでも100以上の項目が考えられるが、本研究では単位が温度の物性(ガラス転移温度や融点など)、電気伝導度(単位:S/cm)、溶解度パラメータ(J/cm³)^{1/2}およびガス透過係数(cm³ cm/cm² s Pa)を対象とした。プログラム言語はPython 3を使用した。

2-2. 評価用データの作成

数値クリーニング、単位統一および数値変換に関する正解データをあらかじめ作成しておき、それぞれの評価で用いた。作成した正解データ数は、物性によって異なるが、170~800とした。

2-3. 数値クリーニング、単位統一、数値変換

数値クリーニングで”±0.01”などの誤差は除去するようにした。また、”110-120”のように複数の数値が記載されている場合は一番左を残すようにした。10 の累乗数についてはその認識を正規表現で行い、数値に反映させるようにした。その他、数字以外の文字はなるべく除去するようにした。単位統一は図1のような基本単位リストをあらかじめ人手作成しておいて行った。図1は電気伝導

度用のリストで、(1)および(2)はそれぞれジーメンズと長さの単位であり、これらの間にはスペースや

```
(0) aida
999 → (no .space)
999.999 → (space)
. → (period)
. → (middle .dot)
(1) siemens (1/ohm)
S → 1
s → 1
mS → 1e-3
mS → 1e-3
μS → 1e-6
μS → 1e-6
Ω-1 → 1
(2) length
/m → 1e-2
m-1 → 1e-2
/cm → 1
cm-1 → 1
```

図1 基本単位リスト。”999”

はプログラム内で消去。

ピリオドなどが入るため、(0)にそのような文字を登録しておいた。このリストを用いて、(0)~(2)の文字をプログラム内で組み合わせることで、”S/m”や”S. m-1”などの単位を自動生成させた。物性によっては基本単位の記載順序が不定であり、そのときはすべての順序を考慮した。リストには個々の基本単位の変換係数も登録しておき、例えば、μS/m → S/cm での数値変換では1e-6×1e-2=1e-8で必要な変換係数を自動生成させた。この変換係数を用いて、数値クリーニング後の数値を変換した。統一する単位には PoLyInfo 既定の単位を用いた^[1]。

3. 結果と考察

電気伝導度での数値クリーニング、単位統一および数値変換の精度はそれぞれ 94.8%、99.2%および 94.8%で、いずれも高精度であった。単位統一が 100%に近い値であったため、数値クリーニングの値がそのまま数値変換の精度となった。数値クリーニングで上手くいかなかった例には“(1.94±0.01)×10⁻⁵”があった。単位統一では”μS”のように長さ部分の単位が省略されたものは統一できなかった。

4. 展開

他の物性についても電気伝導度と同様に検討を行った。ガス透過係数など基本単位が3つ以上から構成される物性については、前半と後半部分に分けて行った。発表当日はその詳細や現在検討中の他物性についても報告を行う。

[1] PoLyInfo, <https://polymer.nims.go.jp/>

[2] H.Oka et al., “Automatic extraction of polymer data from tables in xml”, Third International Workshop on SCientific DOcument Analysis (SCIDOCA2018), 慶応義塾大学(日吉, 横浜市), 2018年11月12-13日。