2G-6

書誌リンケージに基づく研究分野マッピングの精度検証

蔵川　圭[†]　孫　媛[‡]　相澤彰子[*]

国立情報学研究所[†‡*]

1. はじめに

研究評価の一手法としてインパクトファクターなどの論文引用に基づくビブリオメトリクスがある．評価分析においては，メトリクスを多様な評価軸で区切ることが多く，研究機関，研究分野，出版年などの観点から論じることが多い．中でも研究分野は，専門家によって多様な分類が目的に応じて考案され，時代とともに変化するような複雑多岐にわたる性質がある．

本研究では，多様な研究分野分類によるメトリクスの分析が可能となるように，複数の学術データベースで用いられている研究分野を相互にマッピングすることを試みる．具体的には，科研費データベース KAKEN とトムソン・ロイター社の引用書誌データベースである Web of Science(WoS)の研究分野を対象に，機械学習を用いた書誌レコードリンケージによって，研究分野同士のマッピングを行う[1]．

本報は，このマッピングの精度検証の基礎となる，書誌の同一性判定を行うレコードリンケージの精度について検証を行ったことを報告する．

2. 書誌レコードリンケージ

書誌レコードの同一性判定に我々が適用したレコードリンケージの技術は，

1. i-Linkage[2]を用いて，ソースとなる書誌レコード１件に対しターゲットとなる書誌レコード集合から候補を数件取得し（ブロッキングという），
2. 候補となった書誌ペアを SVM を用いて同一性判定する，

という 2 段階のプロセスからなり，大規模な書誌レコード集合に対し計算コストの最小化と精度の最大化を実現する同一性判定器となっている．

3. 精度検証

3.1. 書誌レコード

書誌レコードリンケージの精度を検証するために，KAKEN に収録された 2009 年度実績報告書に列挙される 105,458 件の文献書誌と，該当候補となる WoS の DBYear2009,2010 の論文書誌 3,843,107 件を用いた．KAKEN の文献書誌と WoS の論文書誌はそれぞれ図 1，図 2 に示す形式で XML によって記述されている．

```
<grant_award id="15104002">
....
 <field code="4201">天文学</field>
....
 <publications>
    <journal_article>
     <author><![CDATA[H.Kawai, et al.]]></author>
     <title><![CDATA[Measurement of Ultra-high Energy Cosmic Rays by Telescope Array(TA)]]></title>
     <journal_title_vol_no><![CDATA[J.Phys.Soc.Jpn.Supplement A 78]]></journal_title_vol_no>
     <pages>108-113</pages>
     <year>2009</year>
     <review>1</review>
    </journal_article>
....
</publications>
....
</grant_award>
```

図 1 KAKEN の文献書誌と科研費細目

```
<REC>
<issue recid="175215165" coverdate="200812" sortkey="3070202595" dbyear="2009">
....
<subjects count="1">
<subject code="BU" edition="SCI">ASTRONOMY &amp;amp; ASTROPHYSICS</subject>
</subjects>
....
<item issue="175215165" recid="175215166" coverdate="200812" sortkey="3070202594" refkey="6481967" dbyear="2009">
....
<source_title>JOURNAL OF GEOPHYSICAL RESEARCH-SPACE PHYSICS</source_title>
<item_title>Ways in which ICME sheaths differ from magnetosheaths</item_title>
<bib_pages begin="" end="" pages="10">-</bib_pages>
<bib_issue year="2008" vol="113"/>
<authors count="2">
<primaryauthor>Siscoe, G</primaryauthor>
....
</REC>
```

図 2 WoS の論文書誌と研究分野

3.2. i-Linkage によるブロッキングの性能

KAKEN の文献書誌をソース，該当候補となる WoS の論文書誌をターゲットとして上位 5 件を取得しブロッキングを行った．i-Linakge は，候補となるターゲット書誌をスコアを付与して返却する．Top-N とは，スコアが降順上位 N 件までのレコードを指しており，Top-5 は上位 5 件のレコード集合である．Top-5 として実際に取得したのは 518,605 件である．ただし，いくつかのレコード集合は付与したスコアによって 5 件未満となっている．

ブロッキングの性能を検証するために，返却されたレコード集合を Top-1 のレコードのスコアでランク付けし降順にソートする．105,458 ランクを 1/200 で刻んでレコード集合に含まれるレコードのスコアをランクごとにプロットしたのが図 3 である．プロットの形状は，書誌ペアの人手判定結果を示している．WoS の書誌レコードは原理的に重複がないので，Top-5 の書誌レコード集合には高々1 件の正解ペアがあり，図では Top-1 に正解が集中していることがわかる．また，

Evaluating the accuracy of a mapping between two different sets of subject categories based on bibliographic record linkage

† Kei Kurakawa, National Institute of Informatics
‡ Yuan Sun, National Institute of Informatics
* Akiko Aizawa, National Institute of Informatics

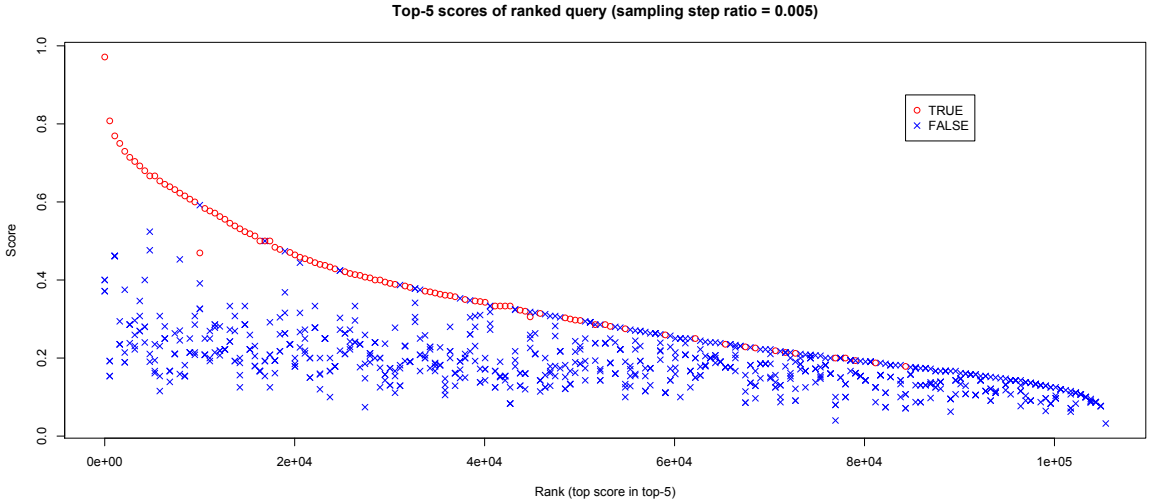

図 3 Top-1 レコードのスコアでランク付けされた Top-5 レコードのスコア分布と書誌ペアの正負判定

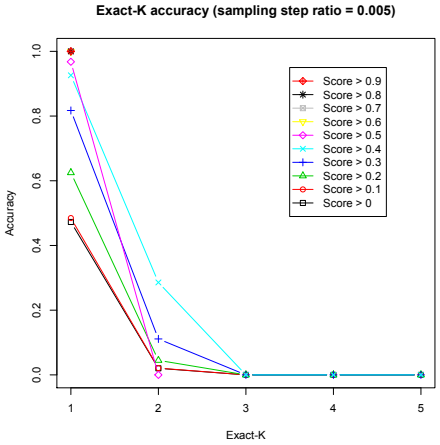

図 4 Exact-K 精度

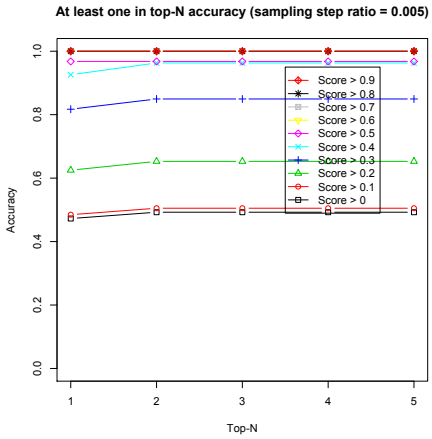

図 5 At Least One 精度

表 1 異なるデータセットごとの SVM の判定精度

Training data set	P / N	Accuracy	Precision	Recall	F-measure
All	99 / 883	**98.98**	**97.00**	93.00	94.89
Top-5 & score>0.28	85 / 84	96.43	96.89	96.53	**96.52**
Top-3 & score>0.27	86 / 83	95.22	95.64	95.42	95.34
Top-1 & score>0.1	95 / 101	96.39	94.27	**99.00**	96.42

Top-1 のスコアが高い場合に正解が現れ，低い場合は正解がなくなっている．これらは該当する書誌が存在すれば高いスコアが付与されるというスコア関数のよい性能を示していることに他ならない．

ブロッキングにおいてスコア上位から何番目に正解があるかを示したのが Exact-K Accuracy（図 4）である．ここでは，上位 K 番目(Exact-K)のレコードの精度をスコアを区切って示している．正解は Exact-1 と Exact-2 に集中していることがわかり，Exact-1 ではスコアの下限を 0 から 1 まで変化させることで，精度が 0.48 から 1.0 まで変化することがわかる．

Top-N のレコードに少なくとも 1 件正解があるかどうかを示したのが At Least One Accuracy(図 5)である．どのスコア下限においても Top-N を変化させたときに Top-1 から精度が増加しないため，Top-2 から Top-5 において正解がほとんど存在していないことがわかる．

3.3. SVM による分類の精度

ここに用いた人手判定の正解データを再利用して Top-N とスコアの下限を変化させて訓練データセットを複数作成し，文献[1]と同様の特徴ベクトルで SVM を構成し精度を測定した．表 1 に結果を示す．

4. まとめ

書誌レコードリンケージの精度について検証を行った．今後は，より厳密な検証のために統計検定を行う．

参考文献

[1] 蔵川圭，孫媛，中村優文，相澤彰子 〝レコードリンケージに基づく科研費分野-WoS 分野マッピング〞 2013 年度統計関連学会連合大会講演報告集，p.111，詳細論文

[2] 相澤 彰子，高久 雅生，大山 敬三「大規模データベースを利用したリンケージシステムの提案と実装」，DBSJ Letters，Vol.6，No.4 ，pp.17-20 (2008)