2022 年 7 月度（第 395 回）ライフサイエンス分科会

開催日時：2022 年 7 月 8 日（金）14:05～14:35
開催場所：Zoom
参加人数：166 名
内容：第 19 回情報プロフェッショナルシンポジウム INFOPRO2022 ver.
医中誌 Web　バージョンアップ　～「ゆるふわ検索」の検証～

記入者：医学中央雑誌刊行会　豊島　一平太
株式会社エムエムツインズ　田邊　稔
旭化成ファーマ株式会社　廣谷　映子

1.　内容
１）医中誌 Web への「ゆるふわ検索」採用について（医学中央雑誌刊行会　豊島 一平太 氏）
２）「ゆるふわ検索」検証結果（OUG ライフサイエンス分科会より）
３）「ゆるふわ検索」について解説（株式会社エムエムツインズ　田邊 稔 氏）

１）「ゆるふわ検索」について
- 機械学習型エンジンを利用した新しい検索手法として試験的に導入
- 検索ボックスに入力されたテキストを分析し、類似度が高い順に結果を提示する
- 類似度は TF-IDF による文章の特徴ベクトルから算出
- 使用シーン：文献検索や医学用語に不慣れな方や発見的検索が目的の方
- 検索窓は大きめにしており、かなりのバイト数を入力して検索できる

２）「ゆるふわ検索」検証結果
・使用したテキスト：
①医療用医薬品の添付文書とくすりのしおり、
②リーダビリティ研究のテキスト（酒井由紀子. 健康医学情報の伝達におけるリーダビリティ. 樹村房, 2018, , 242p.より p.166 IV-2 図 3 種の実験テキストの冒頭部分）
③クリティカルパス：
静岡県立がんセンター.クリティカルパスについて.〇食道外科＞食道切除術を受けられる方用
https://www.scchr.jp/admission/wp-content/uploads/sites/3/2015/06/shokudougeka001.pdf
- 重複状況、論文タイプ、検索結果の上位と下位について結果を提示した。

３）「ゆるふわ検索」について解説
- ゆるふわ検索エンジンの基本処理フロー
- 単純化したベクトル計算の解説
- 検索テキストの単語種類と出現回数
- 将来的な拡張計画案

2.　質疑

Ｑ１）クエリ文章の特徴量は Tf-Idf とのことだが、クエリ文章のトークンへの切り分けは形態素解析か？
形態素解析の辞書には特殊な用語（医学用語やフレーズなど）も使われているか？
品詞選択などの前処理もおこなっているのか？
→Ａ１）クエリ文章のトークンへの切り分けは形態素解析である。
形態素解析の辞書には特殊な辞書は使用していないが、今後の採用は課題として認識している。
品詞選択など、前処理は行っている。（田邊氏）

Ｑ２）テキストマイニングのワードクラウドの文字の大きさは Tf-Idf か？
（成分名が大きくなるのは Idf の重みのためかと推測してのＱ）
→Ａ２）ユーザーローカル社のサイトを利用したため詳細は不明だが、
　スコア順とあるためおそらくご指摘の通りと思われる（ライフサイエンス分科会より廣谷が回答）

Ｑ３）ゆるふわ検索で、関連性の高い文献を出すためのポイントはあるのか
　（文章の文字数の長さ、文章でなくキーワードを並べたようなものがよいか）
→Ａ３）まず、何を以って「関連性が高い」と評価するかは人間と機械で異なりますので、検索のポイントを示すのも難しいところです。その前提の上でコメントさせて頂きます。
検索対象のデータはあくまでも文書（文章）ですので、文脈やニュアンスを考慮するのであれば入力するテキストも文章の方が良いかと思います。一方、特定のキーワードで網羅的に探したい場合は、キーワードやキーワードペアを入力されると良いかと思います。例えば、主たるキーワードと、それとあまり関連のなさそうなキーワードの組合せなど。その際に、主たるキーワードをいくつも羅列されると出現率（TF 値）は高くなりますが、レア度（IDF 値）が下がるため、結果的に主たるキーワードとの関連度（評価）は相対的に下がる傾向があります。
少しでも検索のヒントになれば幸いです。（田邊氏）

Ｑ４）検索対象となるデータは、タイトルと抄録フィールドのみでしょうか。
→Ａ４）検索対象データには、タイトルと抄録に加え、論文種類・収載誌名・著者名・所属機関名・特集名・索引情報（シソーラス語・医中誌フリーキーワード・チェックタグ）を使用している。（医中誌）

3.　次回以降の予定
　9 月 15 日（木）　JST の NBDC について

以上