NAISTにて取り組みたい研究について

氏名: 新妻巧朗
試験区分: 情報科学区分
希望研究室: 自然言語処理研究室

1 はじめに

1.1 NAISTで取り組みたいこと

NAISTにて、私が取り組みたい研究テーマは「情報検索システムにて検索質問に合わせた適切なファセットの生成手法について」である。

このファセットとは、図書館情報学における定義で「あるクラスを2以上の異なる区分特性によって区分したときに得られる下位クラスの総体[1]」のことを言う。また、この区分特性とは「ある分類に属する個々のメンバーに共通する性質」のことである。つまり、このファセットを具体的に説明すれば、共通の性質を抽出することで得られたある区分で検索結果を絞り込むための切り口のことである。

2 研究の概要

2.1 背景・社会的意義

現代社会において情報収集をするためには、検索エンジンを利用することは必要不可欠である。しかし、検索エンジンを適切に活用できず、目的の情報に至れない場面も多い。それは多くの検索エンジンの仕組みが、利用者に対して情報検索能力を要求しているからである。これまでも福島らの研究にて情報検索能力は個人差が大きく、能力差によって情報格差が生じていることが調査されてきた[2]。こうした課題を解決することで、情報に辿りつけないことで生じる機会損失を減らすことができるのではないかと考えている。過去に齋藤らによる教育を通して情報検索能力を向上させる研究[3]も存在しているが、本研究ではシステムの拡張によって解決するアプローチを考えていく。

福島らによって言語能力の高さが情報検索能力の高さに関係しているとわかった[2]。つまり、言語能力の高低が情報検索において、情報格差を生み出していると考えられる。そのため、情報の探索過程で言語能力を要求する場面にて利用者の補助をおこなうシステムを提案したい。

2.2 提案内容

そこで、利用者が入力した検索質問に対して適切なファセットを表示することで、検索質問を検索意図に近づけていくシステム（図1）を提案したい。

検索意図とは、人が検索行動をおこなう動機のことである。情報の探索行動は、検索意図から生じる検索質問を検索の動機を満たす文書に近づけるプロセスであると考えられる。そのため、ファセット検索が利用できるのではないかと考えた。ファセット検索とは、検索システムの利用者に検索対象を何らかの側面で絞り込むファセットを提示し、検索対象を絞り込んでいく検索手法である[4]。これは、システムの利用者が検索意図を言語化する行動をシステムが代行していると言える。そのため、検索エンジンが個人の言語能力に依存している問題にアプローチできると考えている。

図 1: システムのイメージ図

3 研究の方法

3.1 従来のファセット検索の課題

ファセット検索の典型的な用例として、Amazon.co.jp[5]の検索結果画面をあげる。ファセット検索は図2の赤枠で囲われたメニューのように、ある分類に関する検索結果をさらに絞り込む切り口を提供する。先に挙げた例では、商

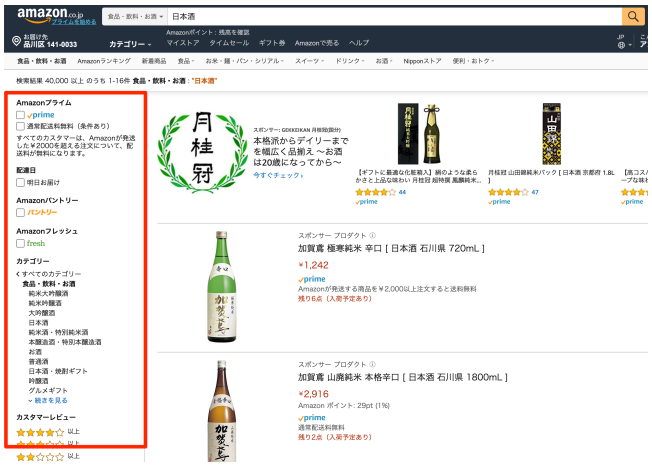

図 2: Amazonの検索結果画面: 日本酒に対するファセット検索

品データを索引対象として扱っていた。このように従来のファセット検索では、索引対象には既に構造化がされており属性データを持った文書を利用することが多い。この属性データとは、ある区分特性に属しているかどうかを示すようなメタデータのことである。

また、ファセット検索を非構造的な文書にて利用する場合には、事前に文書から区分特性を見つけ出して、それを索引可能な形に変換して属性データとして付与することで半構造的な文書にする必要がある。従来であれば、この区分特性を見つけ、属性データの付与する作業は人手を用いておこなわれてきた。しかし、Web を対象にファセットを作成する場合には、文書の増減と文書分類の変化の早さが原因で、人手による作業が現実的でないという課題がある。そのため、区分特性を推測し自動的に属性データを生成することで、ファセット検索を Web にも応用できるようにしたいと考えた。

3.2 研究の方向性

本システムではファセットを作成するために、二つのデータを出力をすることを目標とする。

ファセットを表す tuple のリスト ある語彙 v を上位クラスとすると、v をある区分特性 c を介して関係する語彙の集合 V を下位クラスとして考えるデータからなる $\langle v, c, V\rangle$ 形式の tuple のリスト。

属性データ 文書があるファセットに含まれるかどうかを示す属性データのこと。文書に付与するメタデータとして考える。

3.2.1 ファセットを表す tuple のリスト

入力データを索引対象の文書として、教師なしでこの tuple の情報を抽出する方法を考えたい。区分特性は「関係」をうんちゃらかんちゃらである。そのため、文中の中で語彙の関係を示せる語彙は述語であることから、本システムにおけるは区分特性には述語を利用する。

そのため、区分特性に抽出には OpenIE と呼ばれる研究の成果を活用できる可能性がある。OpenIE は文章から $\langle \arg 1, rel, \arg 2\rangle$ という形式の tuple の形で、情報を抽出をする技術の研究分野である。[6] この文中の述語を上記の rel として、その周辺の語彙などを arg1, arg2 として抽出する。この arg1 を語彙 v と rel を区分特性 c とみなせば、arg1 と rel が一致している tuple が持つ arg2 を下位クラスの語彙の集合 V としてまとめることで、$\langle v, c, V\rangle$ の tuple を作成できると考えている。

例えば、「日本酒」を語彙 v とすると、区分特性 c は「の銘柄は」といった述語となる。そして、その目的語である「一ノ蔵」「花浴陽」などをまとめて、語彙の集合 V を得れば $\langle$"日本酒","の銘柄は",["一ノ蔵","花浴陽",...]$\rangle$ という tuple を作ることができる。

3.2.2 属性データ

3.2.1 にて作成した tuple のうち、上位クラスである語彙 v が文書に関係しているかをブール値で表現するメタデータを文書に付与することで実現できる。tf-idf などに閾値をもうけて関係性を表現することを検討している。

3.2.3 まとめ

3.2.1 の tuple をファセットのユーザインタフェースを作成するのに利用し、検索処理では 3.2.2 で文書に付与した属性データを利用することで、提案をしたシステムが実現できるのではないかと考えている。

4 これまでの修学経験等

学部では地方の産業構造に関する実証分析について研究してきた。特に卒業研究では総生産と地域を構成する産業に着眼し、経済格差が生じる要因について分析をした。また、社会人ではソフトウェアエンジニアとして Web サービスに携わり、検索システムの利用者が得たい情報をどう探索しているのかについて考えてきた。特に現在携わっているアルバイト求人のデータベースメディアでは、どのようにファセットナビゲーションを実現するとよいか、求人検索機能のファセット検索をどのように実装すべきかなどを試行錯誤する機会に恵まれた。こうした経験が本研究では役立つのではないかと考えている。

5 最後に

ここまで NAIST にて取り組みたい研究テーマや自身の経験について述べてきた。私が NAIST を志望するのは、異なるバックグラウンドを持った人間を受け入れるサポート体制が整っており、かつ優れた研究成果を出している大学院であるからだ。こうした NAIST の整った教育・研究環境を活かして、自然言語処理や情報検索の分野に貢献していきたいと考えている。

参考文献

[1] 日本図書館情報学会用語辞典編集委員会編 (2013), 図書館情報学用語辞典 第 4 版

[2] 福島健介・小原 格・須原慎太郎・ほか (2005), インターネット検索能力の差異に及ぼす 要因の検討 その 1, コンピュータ＆エデュケーション VOL.18 2005

[3] 齋藤ひとみ・三輪和久 (2004), Web 情報検索におけるリフレクションの支援, 人工知能学会論文誌 19 巻 4 号 C (2004 年)

[4] Daniel Tunkelang (2009), Faceted Search (Synthesis Lectures on Information Concepts, Retrieval, and Services), pp. 21—26

[5] Amazon.co.jp (最終閲覧日: 2019 年 5 月 23 日), https://www.amazon.co.jp/

[6] Christina Niklaus, Matthias Cetto, Andre Freitas, and Siegfried Handschu (2018), A Survey on Open Information Extraction, Proceedings of the 27th International Conference on Computational Linguistics