その木何の樹？気になる木
―ウェブ情報を利用した気になる情報発見ツール―

1. 背景

近年、ウェブ上に無料で利用できる巨大なデータベースを持つサービスが存在している。例えば、動画共有サイトの YouTube や画像データを扱うFlickr、ウェブ上での百科事典を実現するウィキペディア(Wikipedia)があげられる。なかでも、ウィキペディアなどのサイトでは、膨大な知識情報が惜しみもなく利用可能である。また、その情報を容易に利用できる仕組みが提供されており、サービス同士の情報や機能を連携させた新しいサービスが既に存在している。このように、非常に膨大な情報を自由に扱えるようになっている。

一方で、我々が普段の生活で直面する問題に対して、解決案を考える際などには、未だ手元の限られた情報から判断を下すことが多い。つまり、せっかく問題の大局を包含するような膨大な情報があり、容易に手に入るにも関わらず、様々な問題に対して局所的な情報のみで対処しているといえる。これは、あまりに膨大な情報を一度に扱おうとすると混乱してしまうことや、そもそも多くの情報を上手に扱うことは容易でない、といったことが理由としてあげられる。そのため、ウェブ上の情報をうまく扱える方法が求められている。

2. 目的

本プロジェクトでは、ウェブ上に公開されている巨大なデータベースを上手く利用し、我々が直面する多くの場面で、意思決定や発想支援を行うシステムの構築を目的とした。ウェブ上に存在する膨大な情報の中から必要な情報群を取得し、それらを構造化することで、比較的効率的にウェブ上の知識の力を借り、対象としている問題の解決や意思決定の手助けを行う。

3. 開発の内容

本プロジェクトで開発を行うシステム（以下、本システム）を利用する際の流れを図1に示す。本システムでは、ウェブ上の情報を取得し、ルール付けにより情報間を構造化し、それをユーザに対して可視化をすることで意思決定支援を行う。

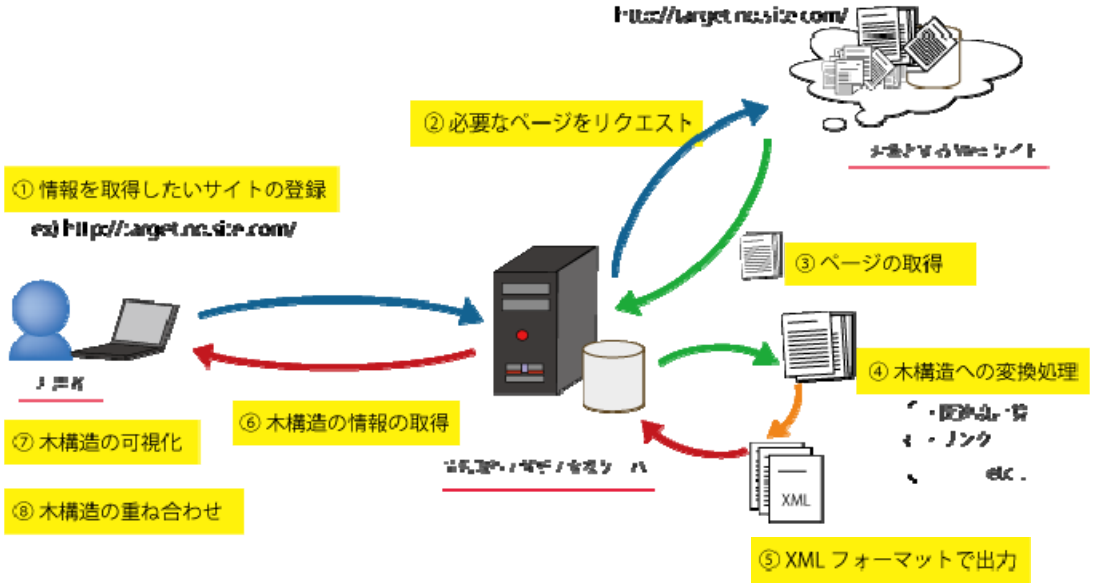

図1　システム全体の流れ

事前準備として、ユーザに対する認証を行う。

① 構造的な情報を利用する際に、ユーザが予め対象としたいサイトのURL、および対象としている分野などのキーワードを登録する。

② システムは、必要なページを取得するために指定されたサイトにリクエストを送る。

③ 必要なページを取得し、取得したページから必要なテキスト情報などを抜き出す。

④ 取得したページの情報に対して、ページ間の類似度計算などを用いて木構造へ変換を行う。

⑤ 木構造に変換した情報を XML フォーマットで出力する。

⑥ ユーザは、木構造の情報を取り出す。このとき、サイトとキーワードの登録の際に発行されたキーを元に、後から結果を確認することもできる。

⑦ 取得した木構造の可視化を行う。

⑧ 可視化を行った後に、可視化を行った木構造に対して、木構造同士の類似度から重ね合わせを行う。

上記の処理の流れから分かるように、本システムでは、大きく分けて情報の取得、情報の加工（構造化）、可視化の3つの処理がある。

3．1　情報の取得

対象サイト内おいてキーワードに関連のあるページを取得するが、まず、その取得対象とするページ群のURLリストを生成する必要がある。そこで、本システムでは、取得リストの生成にウェブ上の検索サービスを利用することとした。

図2に示すように、プログラム内から検索サービスに指定サイト内のページの中からキーワードに関連する、もしくはキーワードを含むページを返してくれるようなクエリのリクエストを送信し、検索結果からリストを取得する。

図2　リストの取得

このリストにより、取得すべきウェブページなどが特定できるので、次に対象とするページの情報を取得する。

3．2　情報の加工

取得した情報を加工して、構造的に組み立てる。様々な方法が考えられるがプロジェクト期間中には、一例として文書の内容の類似度に基づいた構造化を目指した。まず、取得ページから文章情報を取り出し、類似度を計算するための単語の出現頻度テーブルを生成する（図3）。そして、図4に示すように、頻度テーブルにもとづいて類似文書を紐付け、文書間の関係を構造化する。構造化を行った情報はクライアントアプリケーションからのリクエストがあると XML 形式で出力する。

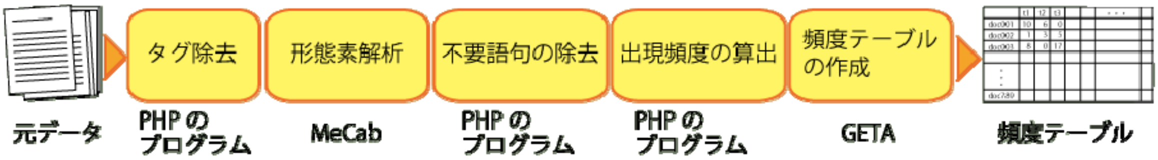

図3　単語出頻度テーブルの生成

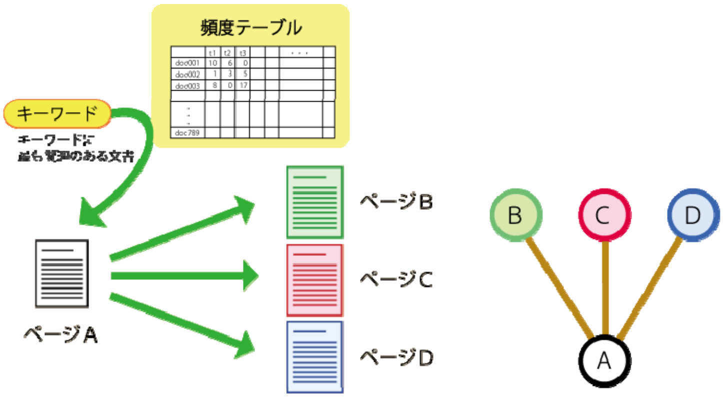

図4　木構造の生成の流れ

3．3　情報の可視化

ユーザはクライアントアプリケーションを通して構造化した情報を活用することができる。効果的な表示によりユーザの発想支援を行う。構築したクライアントアプリケーション上では、木構造にちなんで実際の木のように、構造化した情報を樹で表現している（図5）。

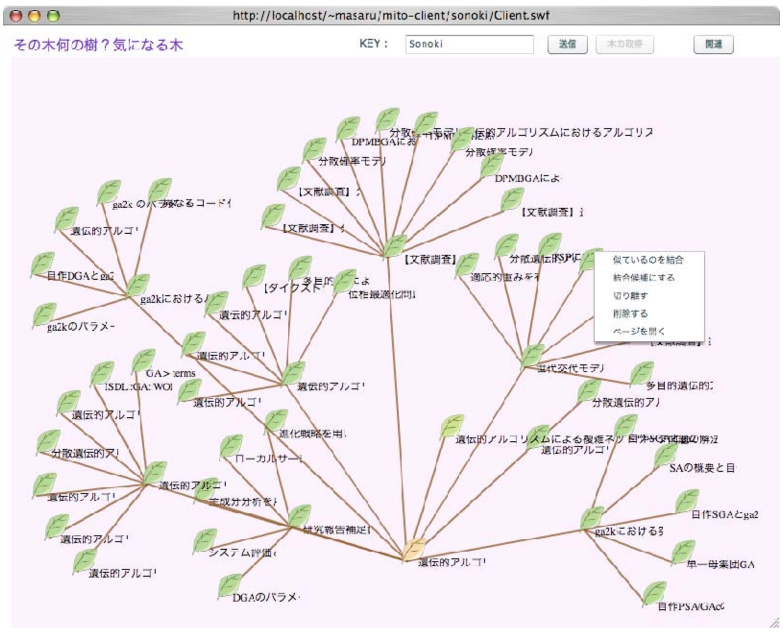

図5　クライアントアプリケーションの実行例

可視化をおこなうことで、それまでに気づかなかった情報を見つけ、発想支援につながると考えられる。また、図6に示すように2つ以上の樹を、それぞれの情報が持つ類似度により、自動的に重ね合わせることができる。

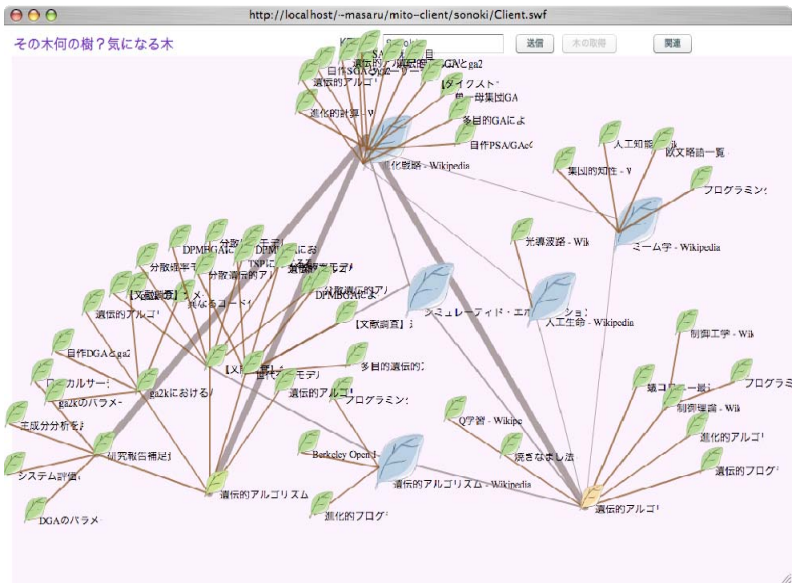

図12　クライアントアプリケーションでの重ね合わせ

樹の重ね合わせをすることで、2つ以上の樹を比較することができ、重なりが多い部分は「重要な情報が集約している」、逆に重なりの少ない部分は、「重要度が低い、もしくは、逆に普段気づかないアイデアにつながる情報である」ということが分かるため、発想が刺激されることや、意思決定を行う上での判断材料にすることができる。このように、情報の構造化と可視化、そして、重ね合わせによる重複と差分を提示することで新たな発想の支援や意思決定の支援が期待できる。

5. 特徴と今後期待される効果

本プロジェクトでは、ウェブ上の情報に対して情報の構造化と可視化を行うというアプローチで活用の方法を模索した。また、複数の構造化した情報から差分をとることで新たな発想や意思決定の材料になりうる情報を引き出す試みが特徴的である。意思決定や発想支援だけでなく、実際の利用例として特許情報や研究レポートなどを対象として利用すると、関連のある特許情報や関連のある研究が視覚的に提示され、また、重ね合わせにより重要な分野や未開拓分野の特許や研究のテーマを発見できることも期待できる。

今後も増大するウェブ上の情報を活用する方針として、ユーザに分かりやすくまとめ、提示することは非常に重要であると考えられる。この点で本システムは、意思決定や発想支援といった枠組みに捕われず、情報をまとめるツールとしての利用も期待される。

6. 普及（または活用）の見通し

近年のウェブ上の情報の増加は非常にめざましく、今後その蓄積された情報をより効率的に利用していくことが、ますます重要となってくることが考えられる。その点で、本システムは、情報をユーザの利用対象に適した加工を行い、可視化を行うフレームワークとしての役割を果たす。システム自体は基本的にフリーの開発環境を利用しているため、導入コストが非常に少ない。今後、オープンソースで公開を予定しており、様々な形で利用されることを期待している。

7. 開発者名（所属）

柴田　優（同志社大学大学院　工学研究科）