# 言語横断型の潜在関係検索エンジンの開発

## —「ドイツの富士山と言えば?」を答える検索エンジン—

### 1. 背景

WWW 空間における情報爆発に伴い、単純なキーワードを含む Web ページ検索だけではなく、エンティティ(人名、組織名など)間の関係に着目した関係検索も有効な場面が考えられるようになってきた。既存のキーワードベース Web 検索エンジンでは、入力されたキーワードを含む文書を見つけ出せるが、エンティティ間の意味関係を利用する検索ができていない。このため、キーワードを知らない場合や良いキーワードを考え出せない場合、必要な情報を検索できない。例えば、Berlin における秋葉原のようなところへ行きたい時に、{(東京, 秋葉原), (Berlin, ?)}のクエリを思い浮かべやすいが、既存の検索エンジンでは答えを簡単に見つけ出せない。

### 2. 目的

本プロジェクトでは、「Berlin の秋葉原と言えば何?」というような質問に対して、答えを直接出す検索エンジンを実現する。即ち、{(東京, 秋葉原), (Berlin, ?)}のようなクエリに対して、潜在関係を根拠にした、はてな(?)に対する適切な地名等を検索できる検索エンジンであり、これを「潜在関係検索エンジン」という。更に、異なる言語に書かれている Web ページも検索できるように、言語横断型の潜在関係検索エンジンを開発する。

### 3. 開発の内容

#### 3. 1. 動作環境

本プロジェクトで開発した検索エンジンは Web 上に書かれている文書をデータとして、動作する。また、検索エンジンはサーバで動作し、サービスとして一般公開している。

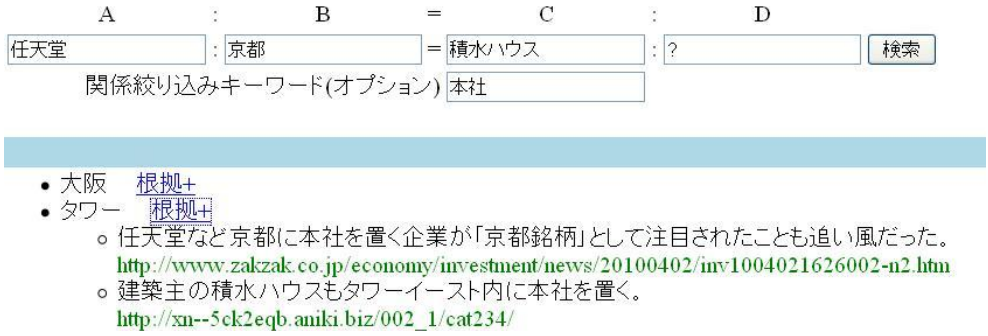


図 1:検索エンジンの動作例

図 1 に検索エンジンの動作例を示す。検索結果は探したいエンティティそのものであり、「根拠」の部分をクリックすると、検索エンジンが Web 上のどんな文を根拠としてその結果を出力したのかが分かる。

3. 2. 検索エンジンの構成

図 2 に検索エンジンの構成を示す。本検索エンジンでは、HBase を使い、索引を保存する。これによりデータベースが分散でき、大量のデータを高速に保存、処理できる。

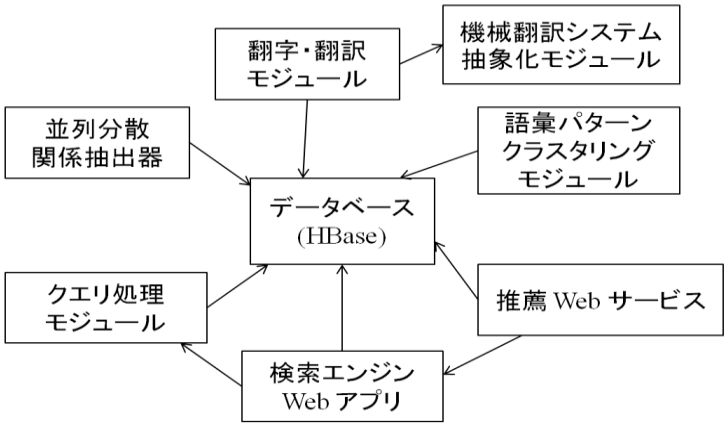


図 2:言語横断型の潜在関係検索エンジンの構成

図 2 における各コンポーネントが独立して動作でき、更に、複製できる。これによって、例えば、「検索エンジンの Web アプリ」を 2 つに複製して、大量クエリに対応できる。

3. 3. 主な機能

3. 3. 1. 潜在関係検索、言語横断型の潜在関係検索

{(東京, 日本), (?,ベトナム)}のような潜在関係検索のクエリ(質問)が入力されたときに、正確に「ハノイ」や「ホーチミン」をトップにランキングする。また、{(東京, 日本) , (?, Vietnam)}のような言語横断型のクエリも処理できる。根拠となる文も取得し、ユーザに提示する。

3. 3. 2. 入力エンティティの推薦機能

図 3 に示すように、ユーザがエンティティを入力するときに、検索エンジンの索引から該当のエンティティを推薦する。1 つのエンティティが入力された場合、そのエンティティとペアをなすエンティティだけを推薦対象にする。

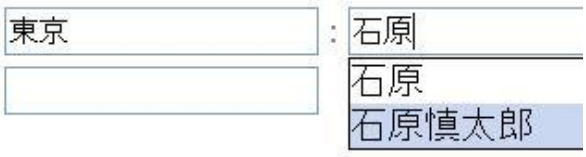


図 3:入力推薦機能

3. 3. 3. 関係絞り込み機能

例として入力するエンティティペアに複数の意味関係があるときに、どの関係を検索したいかを指定できるようにする。例えば、図 1 では、企業と本社所在地との関係を検索したいので、「本社」を「関係絞り込みキーワード」の入力前窓に入力する。ただし、この入力がなくても検索できるが、絞り込むことで、より精度が上がる。

4. 従来の技術（または機能）との相違

- 高速でクエリ処理できる。既存の潜在関係検索エンジンは索引を作成せずに、外部のキーワードベース検索エンジンを利用するのでクエリ処理速度が遅い。本プロジェクトの検索エンジンのクエリ応答時間は普通の Web 検索エンジンの応答時間と同等のものであり、通常のユーザの検索セッション内にクエリの処理ができている。
- 単一言語のクエリに対しては、より高精度の結果を出力することができる。この結果は、語彙パターンのクラスタリングを行うことにより、言い換え表現が吸収され、類似意味のパターンと共起する関係類似度の高いエンティティペアを高精度で見つけ出せたからである。潜在関係検索を高速・高精度に実現できたのは本プロジェクトが初めである。
- 言語横断型のクエリに対しても、ある程度正確な答えを出力できている。これまでに、言語横断的に潜在関係検索を行う技術はなかった。
- 本検索エンジンは語彙パターンとエンティティペアが共起した根拠の文もユーザに見せるので、エンティティ間の関係がよく理解できる。また、言語横断型のクエリでは、根拠の文がほぼ対訳になる場合が多いので、ユーザの翻訳作業支援という使い方もできる。

5. 期待される効果

本検索エンジンはさまざまな分野に応用できると考えられる。

まず、Web 検索エンジンのユーザに対して、英語-日本語の言語横断型の潜在関係検索エンジンを提供することで、色々な使い道があると考えられる。

第一の使い道として、検索対象のキーワードを知らない時にもその対象の情報を検索できる。例えば、Sony の PlayStation ユーザが任天堂の類似する製品を検索したい時に、{(Sony, PlayStation), (任天堂, ?)}を検索すれば、結果として、キーワード「Wii」や「ニンテンドーDS」などが得られる。次に、これらのキーワードをグーグルなどのキーワードベース検索エンジンに入れて、関連情報を検索することができる。

第二の使い道として、簡単な質問応答型のクエリを問い合わせできる。例えば、「Oracle 社の社長は誰か」などの質問は、{(トヨタ, 豊田章男), (Oracle, ?)}などのクエリで答えが得られる。

第三の使い方は、海外観光の時に、日本にある観光地に類似したものや名物などを検索する。例えば、クエリ{(富士山, 日本), (Germany, ?)}でドイツで最も高い山を検索することができる。

また、本検索エンジンは異なる分野、異なる言語間の検索ができるので、分野に跨った特許検索に応用できる。特に、関連した分野間の特許検索を行う際に、ユーザは目的分野のキ

ーワードが良く知らないことが多い。例えば、検索エンジンの専門家がオペレーティングシステムの分野の関連特許を調べたい時に、{(検索エンジン, 索引), (Operating System, ?)}などのクエリで OS 分野の重要な技術、キーワードを検索できる。

その他にも、例えば、名物検索、有名メーカ検索、製品検索などの応用が考えられる。

6. 普及(または活用)の見通し

本検索エンジンの技術を 2011 年 8 月にグーグル社(米国)にプレゼンテーションし、「面白い」という評価を受けたので、今後グーグル社との交渉により、グーグル検索エンジンの一部として利用されるよう働きかけを継続する。これにより、本プロジェクトで開発された技術が世界レベルで使われることを目指す。

また、本検索エンジンをまず政府関係のデータに集中して索引を作成する予定であり、出来上がったときに、経済産業省のオープンガバメントラボ(http://openlabs.go.jp/)で公開できるよう進める予定である。

更に、本検索エンジンは Web 上に一般に公開されているので、だれでも利用できる。索引のサイズを大量に増やすことで、多数のユーザ利用が期待される。なぜなら、本検索エンジンは製品メーカ検索、地名検索、名物検索や自然言語処理、リコメンダーシステムなどのさまざまな分野での応用が考えられるからである。ユーザ数の大きな潜在関係検索エンジンのサイトができる可能性があるので、この新しい検索パラダイムが次世代の検索方法になることができると期待している。

7. クリエータ名(所属):

- グェン トアン ドゥク(東京大学大学院情報理工学系研究科創造情報学専攻 学生)
- ボレガラ ダヌシカ(東京大学大学院情報理工学系研究科電子情報学専攻 助教)

(参考)関連URL:

- http://www.milresh.com 本プロジェクトで開発された潜在関係検索エンジン
- http://www.milresh.com/duc チーフクリエータのホームページ
- http://www.iba.t.u-tokyo.ac.jp/~danushka コクリエータのホームページ