挟みこむ検索：明示的に与えられた観点に基づく補間エンティティの発見

旭　直人†　山本　岳洋†‡　中村　聡史†　田中　克己†
京都大学大学院情報学研究科社会情報学専攻†
日本学術振興会特別研究員 DC1‡

1. はじめに

我々はこれまで，“桶狭間の戦い”と“本能寺の変”の間に起こった出来事は何か，2つの本の間の難しさをもつような本はないのか，といったようなユーザにとって既知の2つのものや出来事，プロセスなどの間にあたる何かを発見するという欲求に注目し，ユーザが入力した2つのエンティティの間にあたる補間エンティティを発見する手法を提案してきた[1]．しかし，これまでの手法ではユーザが入力した2つのエンティティがどのような観点で比較されるべきものか考慮できていなかった．そこで本稿では，ユーザがエンティティ間の観点を明示的に与え，候補語を取得する際にコンテキストを限定することで精度の向上を目指す．また，クラスタリングを行うことで，明示的に観点を与えられなくとも，結果を観点ごとに分けることを目指す．

2. 手法

本稿では，[2]のように予め候補語を用意するのではなく，検索エンジンを用いて補間エンティティを発見する手法を提案する．ユーザが入力した2つのエンティティと補間エンティティについてページ作成者がページで言及する場合，補間エンティティをその2つのエンティティの間に記述すると仮定した．また，aとbの間の補間エンティティを求めたいような場合，文章中でaの前やbの後ろでも頻繁に現れるような語は，そのトピックでの一般的な語である可能性が高い．そこでそうした語は適切な補間エンティティではないという仮定をおいた．

2.1. 候補語集合の取得

システムは2つのエンティティ名 a, b と背景語 p をユーザからの入力として受け取る．背景語 p はコンテキストを絞るための語であり，必ずしも必要ではない．次に，システムはクエリ“$(p\wedge)a\wedge b$”を作り，検索エンジンのAPIを利用して検索結果を取得する．システムは検索結果のスニペットから a と b の間に出現する語を補間エンティティの候補語として収集する．ここで，候補語は2文字以上の漢字及びカタカナ，英数字を正規表現によって抽出した．以降，収集した候補語集合を$C=\{t_1,t_2,\ldots t_n\}$とする．

2.2. 語の出現頻度と出現位置によるランキング

本章冒頭で述べた仮説に基づき a と b の間での語の出現頻度と，aの前やbの後ろで語がどの程度出現しているかという尺度を用いて候補語のランキングを行う．まずは C に含まれるそれぞれの語 t に対し，以下の出現位置による尺度 $F_{a,b}(t)$ を求める．

$$F_{a,b}(t)=\frac{tf_{bet(a,b)}(t)}{tf(t)}$$

$$0<F_{a,b}(t)\le 1$$

$tf_{bet(a,b)}(t)$ は a と b の間での t の出現頻度で，$tf(t)$ は t のスニペット全体における出現頻度である．aの前やbの後にtが多く出現すればするほど，$F_{a,b}(t)$の値は小さくなる．tがaとbの間にだけ現れる場合，$F_{a,b}(t)$の値は最大値 1.0 をとる．$F_{a,b}(t)$と$tf_{bet(a,b)}(t)$を用いて以下のようにそれぞれの語のスコアを求める．

$$Rank(t)=\alpha\cdot\frac{tf_{bet(a,b)}(t)}{\sum_{t'\in C}tf_{bet(a,b)}(t')}+\beta\cdot F_{a,b}(t)$$

ただし，

$$\alpha+\beta=1,\quad 0\le\alpha\le 1,\quad 0\le\beta\le 1$$

$$0<Rank(t)\le 1$$

である．システムはこの $Rank(t)$の降順により，候補語のランキングを行い，ユーザに提示する．

2.3. クラスタリング

コンテキストの違いにより，補間エンティティは異なると考えられるので，これをクラスタリングにより分類する．まず，スニペットごと

Interpolation Search: Finding Interpolated Entities by Explicit Aspect Terms from the Web
†Department of Social Informatics, Graduate School of Informatics, Kyoto University
‡Research Fellow of the Japan Society for the Promotion of Science (DC1)

に a と b の間の記述だけを求め，そこに含まれる単語からそのスニペットの特徴ベクトルを作成する．ここで，a と b の間の記述がないスニペットに関しては除外する．このようにして求めた特徴ベクトルから k-means 法により，スニペットのクラスタリングを行う．クラスタリングを行った後，それぞれのクラスタで先に述べた候補語の取得及びランキングを行う．また，それぞれのクラスタのラベルとして，a の前や b の後に出現した語のうち，その語が含まれているスニペット数が高い語を表示する．

3. 適用例

従来手法に背景語を加えることで特定のコンテキストにおける結果を得ることができるのか，クラスタリングによって，コンテキストを分離することができるのかを調べるため，いくつかの例で実験を行った．実験における検索結果取得件数は 200 件である．

クエリ例 1：ワカナゴ→?←ブリ

クエリとして出世魚である“ワカナゴ，ブリ”を与えると，結果は上位から“イナダ”，“ワラサ”，“ヤズ”，“ハマチ”，“メジロ”が抽出された．この結果は関東での呼び方と九州での呼び方が混ざっている．ここで，クエリに背景語として”関東”を与えると，“ワラサ”，“イナダ”が上位にきた．背景語として“九州”を与えた場合，“ヤズ”，“ハマチ”，“メジロ”が上位にきた．背景語が有効に働いていることが確認できた．

“ワカナゴ，ブリ”の結果に対してクラスタリングを k=4 で適用した場合，九州の結果が固まったクラスタと関東の結果が固まったクラスタ，ノイズのみのクラスタ 2 個が生成された．クラスタのラベルも“九州”，“関東”と現れているが，それ以外の語も多く含まれており，“関西”といったような一見すると正しいような語も現れており，ラベル付けは検討の余地がある．また，k-means を用いているため，初期値により結果が変わってしまうため，うまく分離できない場合があった．

クエリ例 2：ナイル→?←黄河

クエリとして“ナイル，黄河”と与えると，結果の 1 位は“インダス川”であった．これは古代四大文明が栄えた土地にある川である，というコンテキストであると考えられる．そこで，四大文明ではなく，川の長さで 2 つの川の間を求めたい，という意図で背景語として“長さ”を加えたところ，1 位は“アマゾン川”になった．これは世界の川の長さランキングというコンテキストで考えた場合である．クラスタリングを行った場合，初期値がうまくいった場合に“アマゾン川”と“インダス川”を分離できていたが，どのクラスタもラベルが同じようなものになってしまっていた．今回は 200 件のスニペットを用いたため，“アマゾン川”が間にくるようなコンテキストをもつスニペットが極端に少なかったため，初期値がうまく設定されないと分離に失敗すると考えられる．クラスタリングがうまく作用するためにはそれぞれのコンテキストを含むスニペットがある程度必要であると考えられる．

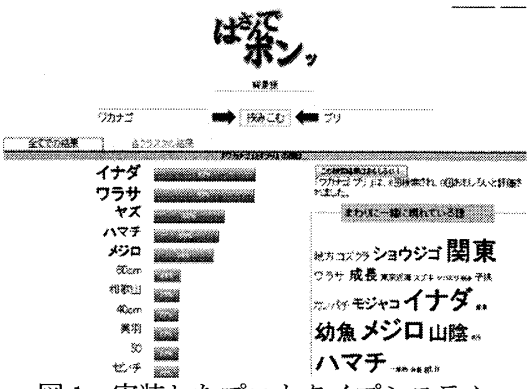

図 1．実装したプロトタイプシステム
(http://hasande.com/)

4. まとめと今後の課題

従来手法に背景語を加えることで意図するコンテキストの結果を得ることやクラスタリングを適用することで，ある程度のコンテキストの分離を行うことが確認できた．しかし，クラスタリングによる分離はまだ十分であるといえず，クラスタのラベル付けの方法に検討の余地がある．また，複数のページにある部分列を集約することや，比較級などで表された順序といったものにも今後取り組んでいく予定である．

5. 謝辞

本研究の一部は，グローバル COE 拠点形成プログラム“知識循環社会のための情報学教育研究拠点”，計画研究“情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究”（研究代表者：田中克己，A01-00-02，課題番号 18049041），未踏 IT 人材発掘・育成事業 2009 年度上期 未踏ユースによるものです．ここに記して謝意を表すものとします．

参考文献

[1] N. Asahi,T.Yamamoto, S. Nakamura, K. Tanaka, “Finding Intemediate Entity between Two Examples on the Web”. Proc. of WIDM 2009, pp. 83-87, 2009.

[2] N. Rubens, V. Sheinman, and T. Tokunaga, “Order Retrieval”, LKR2008, Lecture Notes in Computer Science, vol. 4938, pp. 310-317, 2008.