Webページのアンカーテキストを用いた名前付き実体の分類と検索への応用

吉岡　真治［北海道大学大学院情報科学研究科／准教授］

背景・目的

近年のインターネット経由の情報発信の増加に伴い、今まで手に入れることが困難であった、多くの情報が入手可能になっている。一方で、利用者が適切な検索語を選ばないと、膨大な検索結果が提示され、肝心の欲しい情報を見つけることが難しいという問題がおきている。本研究では、固有名を持つ人物や組織である名前付き実体の検索時における同じ名前（あるいは名前の一部を含む）の別組織といった同表記異実体の区別という問題に注目し、検索語の一部ないしは全部を含む名前付き実体の分類結果を提示することにより、利用者に適切な検索語を選択させる方法を提案することが、本研究の目的である。

内容・方法

本研究では、名前付き実体の分類を行うために、Webページで用いられるアンカーテキストを利用する。アンカーテキストとは、リンク先のWebページを示す説明文であり、特定のWebページが他のWebページの作者からどのような内容のページであると判断されているかを示す情報、あるいは、ページの最も簡単な要約を示した文と考えられるものである。本研究では、同一のWebページをさしているアンカーテキストは、同一の名前付き実体を参照していると仮定し、検索語に対応する名前付き実体とそれに付随する関連Webページを表示することが可能な情報検索システムの構築を目指す。

ただし、アンカーテキストを名前付き実体の名称として利用する際の問題点としては、「戻る」や「トップページ」といった、実体の名前ではないテキストを排除する必要がある。本研究では、アンカーテキストとして利用されるテキストの分類を行う基準を提案すると共に、分類自動化のための検討を行う。

結果・成果

本研究では、まず、最初の予備実験として、日本語Webの大規模テストコレクションであるNTCIR Webテストコレクション NW100g（100GBの主に日本語を中心として集められたWebページのコレクション：15,026,516ページ、リンク数109,133,114）を用いた名前付き実体の検索システムを作成した。具体的には、まず、既存研究でよく利用されているリンクの分類「サイト内リンク（同じホストに対するリンク）」と「サイト外リンク（異なるホストへのリンク）」を利用し、サイト外リンクに対応するアンカーテキストを用いて、Webページを単位とした名前付き実体の検索と分類のシステムを作成した。

本システムでは、名前付き実体の正式名称や略称を入力することにより、対応するページと関連するアンカーテキストを閲覧することができる。また、アンカーテキストに存在するテキストに対し、専門用語抽出のプログラムを適用することにより、よく利用される略称などの情報が抽出可能であることを確認した。

また、本システムの発展形として、Yahoo APIを利用し、名前付き実体に関する検索⇒トップページの候補の発見⇒アンカーテキストによる分析を行うシステムを作成した。例えば、このシステムに対し「ノーステック」という検索キーワードを与えると、ノーステック財団のトップページを含むページを検索される。さらに、各々のページに張られたリンクに含まれるアンカーテキストを利用した専門用語抽出をかけることにより、ノーステック財団のページからは「北海道科学技術総合振興センター」という正式名称が抽出され、その他の、ノーステックテレコム、語学学校紹介のNorth Techとは違う関連語が抽出されることを確認した。これは、日々増加する新語などに対しても、Webを利用することにより対応可能になることを確認したと捉えることができる。

しかし、既存のサイト内かサイト外かといった単純なアンカーテキストの分類では、分析対象となるアンカーテキスト中に名前付き実体の情報を抽出するという観点からは不適切と考えられるデータが存在し、結果として、用語として抽出された語のレベルにばらつきが生じるという問題が発生した。

そのため、本研究では、適切なアンカーテキストのみを用いた情報抽出を行うためのアンカーテキストの分類基準を提案した。この分類では、「リンク先の内容を表すテキスト」、「ページの機能を表すテキスト」、「リンク先との関係を表すテキスト」、「トップページを指示するテキスト」、「ナビゲーションを指示するテキスト」、「インデックスを表すテキスト」、「URLをそのまま利用しているテキスト」、「その他」、の計8種類の分類を提案し、その分類を自動化するための方法論を提案した。ただ、この自動化を行うためには、具体的な事例収集を含む大規模な情報収集と分析が必要となるため、現時点では、完全な自動分類を行う段階ではないが、継続的に事例収集と分析を行っていく予定である。

今後の展望

本研究では、多様なユーザが参加することで日々変化していくWebの情報を用いることにより、多様なユーザの表記ぶれに対応した名前付き実体の検索システムが構築可能であるという可能性を示した。しかし、3.でも述べたように、より精緻な情報を抽出し利用するためには、アンカーテキストの性質に関する更なる分析が必要であると考えている。この分析を進めることにより、提案システムの性能向上を図ると共に、さらなる改良を行い、一般ユーザが利用可能な情報アクセスシステムの構築を目指したいと考えている。