Web情報の俯瞰的把握のための主要・対比・対立文の抽出と集約

河原 大輔 †　　乾 健太郎 †‡　　黒橋 禎夫 †§
dk@nict.go.jp　inui@naist.jp　kuro@i.kyoto-u.ac.jp

†(独) 情報通信研究機構　‡ 東北大学　§ 京都大学

1　はじめに

今日、Web 上の情報は爆発的に増え続けており、さまざまな事柄について多種多様な報道、主張、意見などが存在するようになっている。人々は、その膨大な情報の中から欲しい情報を探し出すために、Yahoo!や Google などの検索エンジンを利用している。すなわち、知りたい事柄をキーワードとして検索エンジンに入力し、そのキーワードに適合する Web ページのリストを得る。そして、この Web ページのリストを眺め、ランク上位の数〜数十件程度のページを実際に見るということを行っている。

しかし、このような情報アクセス手法は、俯瞰性に乏しく、そのクエリについて要はどういうことなのかということを理解することが難しい。また、検索結果上位のページを見るだけでは、そこから得られる情報が偏っている可能性がある。たとえば、アガリクスという食品について既存の検索エンジンを用いて調べてみると、健康に良いという宣伝をするページが上位に表示され、その他の情報は下位のどこかに埋もれてしまうという問題がある。実際には、アガリクスの健康に対する効果が疑わしいことを指摘するページも存在しているが、それを見逃してしまい、一方の偏った情報、意見しか得られない恐れがある。

本研究では、上記の問題を解決するために、与えられたクエリ (トピック) に関する Web 上の情報を集約し、俯瞰的に情報を提示する。提示する情報としては、次の三つとする。

- トピックに関連する主要な文 (主要文)
- 対比されている語・句 (対比キーワード) および文 (対比文)
- 対比キーワードに関して対立している文 (対立文)

たとえば、トピック「合成洗剤」においては、「合成洗剤を使う」「合成洗剤で汚れが落ちる」などが高頻度に出現し、主要な文となっている。対比キーワードとして「合成洗剤」と「石けん」があり、対比文「石けんは合成洗剤と違い、自然に優しく環境にも優しい。」において対比されている。さらに、対比キーワードを含む文では「合成洗剤が環境に悪い」と「石けんが環境に良い」という文が対立している。本研究ではこれらを抽出し、ユーザに提示する手法を提案する。

本手法はこのように、対比・対立関係に着目し、与えられたトピックに関する Web ページ集合中において対比関係にある語や句、さらにはそれらの間で対比・対立している文を抽出、提示する。これは、人々は社会生活を営む上で常に比較や対比を行っているため、物事を理解する際に対比対象を提示した方がより理解が進むと考えられるからである。

また、主要・対比・対立文の抽出は、与えられたトピックに関する Web ページ集合中に出現する多種多様な文に対して、同義、対比、対立関係を同定しながら集約することによって行う。これにより、そのトピックに関してどのような論点があるのかを一覧として見ることができるようになるため、ユーザにとってトピックに対する俯瞰的把握がしやすくなると考えられる。

2　対比キーワードと主要・対比・対立文の抽出と集約

本研究では、与えられたクエリ (トピック) について Web 上にどのような情報があるのかを俯瞰的に把握するために、そのトピックにおける対比キーワードを抽出し、さらに主要・対比・対立文を抽出、提示する。

主要・対立文の単位としては、述語項構造を用いる。述語項構造とはテキスト文書中の「誰が何をどうした」といった文中の単語間の意味的関係であり、これを単位とすることによって、文書分類、要約、意味解析や、既存知識との比較、整合性検証といった論理的分析を

行うことができると考えている。主要・対比・対立文の抽出は、多様な述語項構造に対して、同義、対比、対立関係を同定しながら集約することによって行う。

対比キーワードおよび主要・対比・対立文は次の手順で抽出する。

1. 述語項構造の抽出と集約、および主要文の抽出
2. 対比キーワードと対比文の抽出
3. 対立文の抽出

以下では、まず、関連キーワードと述語項構造の抽出と集約について述べた後、対比キーワードと対比文の抽出、対立文の抽出の方法について詳しく説明する。

2.1 述語項構造の抽出・集約と主要文の抽出

述語項構造の抽出は、形態素・構文解析の結果から、述語と、それに係る一つ以上の項を抽出することによって行う [6]。項としては、文節内の自立語列とし、述語としては、句読点、括弧と一部の機能語 (「ます」「いる」など) 以外の形態素列を抽出し、最後の形態素だけ基本形にする。述語に、モダリティ、否定、目的、条件を表す表現を含んでいれば、それぞれを表すフラグを付加する[1]。

上記の処理を、与えられたトピックに関する Web ページ集合を対象として行う。この Web ページ集合は、トピックをクエリとして検索エンジン基盤 TSUBAKI[4] に入力することによって取得する。取得する Web ページ数は 10,000 件とする。各 Web ページに対して次の処理を行うことによって、キーワード (名詞句) と述語項構造を抽出する。キーワードは後述のキーワード蒸留を行うために抽出する。

1. Web ページから重要文を抽出する。重要文としては、トピックを含む文の周辺とする。抽出する重要文の数は、各ページから 15 文とする。
2. 重要文に対する、形態素解析器 JUMAN[2] と構文解析器 KNP[3] の結果からキーワードおよび述語項構造を抽出する。この処理を高速に行うために、形態素・構文解析を動的に行うのではなく、Web ページ標準フォーマット [4] に格納されている解析結果から取得する。

[1] 本研究の抽出対象は、述語にモダリティなどの態度を付与しているので、「陳述」と言った方が正しいが、本論文では「述語項構造」と呼ぶことにする。

[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

[3] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

このようにして抽出した述語項構造は、項や述語の表現にバリエーションが大きく、同じ意味をもっているが表現上は異なっている述語項構造が散在しているため、これらを集約する。基本的には、同一もしくは同義の述語項構造を一つのクラスタにまとめ、その頻度を計数する。

項に関しては、キーワード蒸留 [5] を用いて、同じ意味を表す項の認識を行う。キーワード蒸留とは与えられた Web 文書集合から、表記揺れや同義表現、部分全体関係のキーワードを徐々に集約していくことによって、トピックに関連するキーワード (関連キーワード) を抽出する手法である。述語に関しても同様に、同義表現の認識を行う。これらの同義表現は、国語辞典や Web から自動獲得した知識を用いる。以下にマージされた述語項構造の例を示す。

(1) a. 石けんを使う
    b. 石鹸を使う
    c. 石鹸を使用する

このようにして集約された述語項構造を頻度順に並べ、頻度上位の述語項構造を主要文とする [6]。

上記のキーワード蒸留手法は、最後に得られるキーワード集合に対して、TF-IDF に基づくスコアを与え、高いスコアのキーワードをトピックの関連キーワードとして出力する。本研究では、この関連キーワードにトピック自体の語・句を加えて、後の処理で用いる。

2.2 対比キーワードと対比文の抽出

対比キーワードは、関連キーワードが対比の構文 (述語項構造) で用いられている場合に抽出する。関連キーワードに限定するのは、トピックに関連しているキーワード間のみの対比関係を抽出するためである。対比キーワードの抽出手順を以下に示す。

1. 「X は Y と (は) 違う | 異なる | 逆だ | 比べる」のパターンにマッチする述語項構造を抽出する。
2. X と Y の両方が関連キーワードに含まれているならば、(X, Y) を対比キーワードのペアとして抽出する。

対比文は、上記 1 にマッチした述語項構造を含む文全体として、これを抽出する。

たとえば、次の文の下線部から、(石けん, 合成洗剤) という対比キーワードを抽出し、この文全体を対比文として抽出する。

(2) 石けんは合成洗剤と違い、自然に優しく環境にも優しい。

2.3 対立文の抽出

対立文は、項が対比関係、述語が対立関係にある述語項構造のペアと定義し、これを述語項構造の集約結果から抽出する[4]。まず、項が対比キーワードのいずれかである述語項構造を抽出し、対比キーワードごとに整理する。対比キーワードペア間において、述語が対立している文があれば対立文として認識する。述語の対立としては次の二種類を考える。

- 否定
  述語に否定フラグがついていなければ、否定フラグを付加した述語項構造、逆に否定フラグがついていれば、それを削除した述語項構造が存在する場合に対比文として認識する。

  (3) a. 石けんが環境に 良い
  b. 合成洗剤が環境に 良くない

- 反義語
  述語を反義語に置換した述語項構造が存在すれば、対比文として認識する。反義語への置換は、国語辞典から抽出された反義語辞書 [3] を利用する。

  (4) a. 石けんが 安全だ
  b. 合成洗剤が 危険だ

3 実験

「合成洗剤」「海洋深層水」「カテキン」「電子マネー」「裁判員制度」「特許制度」「Windows Vista」などの 50 個のトピックについて、対比キーワードおよび対比・対立文の抽出実験を行った。主要文の抽出についての実験、評価は [6] を参照されたい。

3.1 対比キーワードと対比文の抽出結果

50 トピック中 25 トピックについて対比キーワードおよび対比文が抽出された。25 トピックに対して合計 54 ペアの対比キーワードが得られたので、1 トピックあたり約 2 ペアの対比キーワードが抽出された。抽出された対比キーワードの例を表 1 に示す。トピック語自体との対比キーワードとともに、トピック語以外で対比されているキーワードも抽出されている。

表 1: 抽出された対比キーワードの例

・トピック: 合成洗剤
(合成洗剤, 石けん) (液体石けん, 粉石けん)
・トピック: 裁判員制度
(裁判員制度, 陪審制度) (裁判員制度, 参審制)
・トピック: 特許制度
(特許制度, 実用新案制度) (大企業, 中小企業)
・トピック: Windows Vista
(Vista, XP)

抽出された対比キーワードを人手で評価したところ、約 74%(40/54) が実際に対比されて使用されていた。主な誤り原因は、対比部分の一部が省略されていることであった。たとえば、トピック「特許制度」において (アメリカの特許制度, 日本) という対比キーワードペアが抽出された。これは、次の対比文の下線部から抽出されたものである。

(5) アメリカの特許制度は日本と異なり、出願された特許はすべて審査が行なわれ、···

この対比文自体は誤りではないが、下線部は「アメリカの特許制度は日本の特許制度と異なり、」が省略されたものであり、この省略が対比キーワードの誤った抽出の原因となっている。この問題に対する対策として、このような省略の補完ができればよいが、もし出来なくても、対比キーワードペア間の類似性によるフィルタリングをすれば、省略表現からの抽出をしないようにすることができると考えられる。

3.2 対立文の抽出結果

対比キーワードが抽出された 25 トピック中 9 トピックにおいて対立文が抽出された。9 トピックに対して抽出された対立文は合計 26 ペアであり、これを人手でチェックしたところ、否定の認識誤りなどもなく、すべてが本研究における定義と合うものであった。抽出された対立文の例を対比文とともに表 2 に挙げる。

このように、実験の対象とした 50 トピック中において、対立文が抽出されたのは 18%の 9 トピックにとどまった。対立文がそれほど抽出されていない原因は、対立として認識する要素が否定と反義のみであり、対立の範囲が厳しすぎたことが挙げられる。今後は、極性 (肯定・否定) 分類の技術を利用し、極性の違いも対立とみなすことによって、対立文の拡充を行いたいと考えている。

[4]本論文における「対立文」は、真に対立、矛盾していないが、述語が対立関係にあることから、このように呼ぶことにする。

表 2: 抽出された対比・対立文の例

トピック: 合成洗剤			
対比文:	石けん は 合成洗剤 と違い、自然に優しく環境にも優しい。		
対立文:	合成洗剤を使わない	↔	石けんを使う
	合成洗剤が環境に良くない	↔	石けんが環境に良い
	合成洗剤は危険だ	↔	石けんが安全だ
	合成洗剤は分解されない	↔	石けんは分解される
トピック: 電子マネー			
対比文:	電子マネー は 現金 と異なり、そのバリューを外見上から判断することができない。		
対立文:	電子マネーを使う	↔	現金を使わない
	電子マネーを使わない	↔	現金を使う
対比文:	Suica は、常時オンラインで使用しなければならないようで、この点が Edy と異なります。		
対立文:	Suica が使えない	↔	Edy が使える
	Suica が使える	↔	Edy が使えない

また、対立文として表 2 のような文が抽出されたが、これを起点としてこれらの文が出現した文を一覧として見ることによって、トピックについての俯瞰的な把握ができるようになる。我々は、これまでに情報分析システム WISDOM[1] を開発しており、その上では主要文の一覧に加え、選択した主要文の KWIC(keyword in context) 表示を行えるようになっている (図 1)。本研究で抽出した対比・対立文に対しても、このような文脈付きの表示ができるようになれば、情報の俯瞰性が高まると考えられる。

4 関連研究

本研究では、述語項構造の同義、対比、対立関係の認識を行ったが、含意、対立関係に関しては、近年、英語を対象として RTE (Recognizing Textual Entailment) と呼ばれる評価型のワークショップが開催されており、活発に研究が行われている [2]。

佐藤らは blog から比較関係を抽出する手法を提案している [7]。この手法は、「東京より大阪は雰囲気が良い」のような比較表現から比較の基準、対象、属性、評価の 4 つ組を抽出するものであり、本研究の対比文よりも幅広い比較現象を扱っている。一方、本研究の対立文は文間の関係であるため、彼らが対象としている比較表現とは異なる。

5 おわりに

本稿では、与えられたトピックを俯瞰的に把握するために、そのトピックにおける対比キーワードおよび主要・対比・対立文を抽出、提示する手法について述べた。対比キーワード、対比・対立文の抽出実験を行い、その有効性を示した。今後は、これを情報分析システム WISDOM に組み込む予定である。WISDOM において、対比キーワードや主要・対比・対立文の一覧を見ることにより、与えられたトピックに関する情報をより多角的に俯瞰することができ、信頼できる情報の把握に役立つと考えられる。

図 1: WISDOM の主要・対立文提示

参考文献

[1] Susumu Akamine, Daisuke Kawahara, Yoshikiyo Kato, Tetsuji Nakagawa, Kentaro Inui, Sadao Kurohashi, and Yutaka Kidawara. WISDOM: A web information credibility analysis system. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pp. 1–4, 2009.

[2] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 1–9, 2007.

[3] Tomohide Shibata, Michitaka Odani, Jun Harashima, Takashi Oonishi, and Sadao Kurohashi. SYNGRAPH: A flexible matching method based on synonymous expression extraction from an ordinary dictionary and a web corpus. In *IJCNLP2008*, pp. 787–792, 2008.

[4] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. TSUBAKI: An open search engine infrastructure for developing new information access methodology. In *IJCNLP2008*, pp. 189–196, 2008.

[5] 馬場康夫, 新里圭司, 柴田知秀, 黒橋禎夫. キーワード蒸留型クラスタリングによる大規模ウェブ情報の俯瞰. 情報処理学会論文誌, Vol. 50, No. 4, pp. 1399–1409, 2009.

[6] 河原大輔, 黒橋禎夫, 乾健太郎. 主要・対立表現の俯瞰的把握 ― ウェブの情報信頼性分析に向けて. 情報処理学会 自然言語処理研究会 2008-NL-186, pp. 49–54, 2008.

[7] 佐藤敏紀, 奥村学. blog からの比較関係抽出. 情報処理学会 自然言語処理研究会 2007-NL-181, pp. 7–14, 2007.