専門用語集を検索インタフェースとする文書群アクセス

2P－4

岩本　元†　　西野　文人‡　　柿元　俊博‡
†関西電力　‡富士通研究所

1.はじめに

インターネットに代表される大規模な文書群から文書を探す場合、全文検索にはキーワードの多義性、複数キーワード間の関係を詳細に指定できない等の問題があり、その適合率は低い。また、探すのは文書そのものでなくそこから抽出される情報である場合も多い。本稿では、文書群から情報抽出により特定専門分野の用語集を作成し、その分野の利用者がその用語集をインタフェースとして文書群にアクセスする方法を検討する。

2.コンセプト

現在、大量の電子文書がインターネットを中心に発生しており、多くの検索サービスが提供されている。そのほとんどはキーワードのマッチングによる全文検索であるが、次のような問題からその適合率が低い。

(1) キーワードの多義性

例えば、「ネットワーク」には「通信用のネットワーク」「テレビ局やラジオ局の番組供給網」等の複数の使い方があるため、「ネットワーク」を条件とする全文検索結果にはそれらの全てが含まれる。

(2) キーワード間の関係を詳細に指定できない

キーワードの多義性の問題はキーワードを複数個指定することである程度解決できる。前述の場合、「ネットワーク and 通信」とすれば意味が限定される。その場合、キーワード間の関係は共起関係の論理積(and)か論理和(or)である。しかし、例えば、「ネットワーク and 通信」にヒットする「通信ネットワーク」と「通信会社社員間のネットワーク」の意味は異なる。また、2つの単語が同じ文の中に現れることは通常の全文検索では保証されない。

(3) 良いキーワードを思いつかない

従来のシソーラス辞書を用れば、キーワードの範囲を変えたり、類義語を得ることができる。しかし、特定の専門分野に対してあらたにシソーラスを構築する手間はかなり大きい。また、文書群から得た単語の共起関係を基にして関連語を提示する方法もあるが、単語間の意味的な類義性は弱い。

そこで、単語間の意味的な関係を文書群から抽出して用語集を作成し、その用語集をインタフェースとして文書群にアクセスする方式を検討した。この方式は先述の問題に対して次のような対策を提供する。

(1) 用語の複数の意味を提示し、利用者が多義性を認識して正確なキーワードを指定する
(2) 用語間の関係を詳細に指定して検索する
(3) 意味的に関係する用語を参照してあらたなキーワードを得る

また、この方式には次のような効果も期待できる。

・最終的に得たい知識が予め抽出されている。
　（詳しく知りたいときは抽出元文書を読む。）
・複数の知識を一覧比較できる。
・参考文献として複数の文書間の関係がわかる。
・文書の校正において用語の使い方を統一する。

3.要求仕様

3.1　用語集へのアクセス

用語集の使い方は次の3つとなる。

①文書検索時にインタフェースとして使う
②読書等の作業時に用語を引く
③文書作成時、作成中の文書から用語を抽出して用語の使用方法等をチェックする

①②のときは下記の用語集アクセス方法が代表的に必要である。

・用語の辞書引き（部分一致）
・用語間の関連リンクを辿る
・関連による用語の逆引き（定義中の全文検索等）

他にも、電子用語集として多角的な情報をユーザの要求に合わせて様々な形で提示することが可能である。

3.2　文書へのアクセス

用語集の各項目から該当する用語の属性（定義、訳語、性質等）、他の用語との関係、抽出元文書へのリンクが張られる。それを辿ることによって文書へアクセスする。

4.用語集の構築

4.1　用語情報

用語情報は、項目名・属性（“定義”・“訳語”・“性質”等）・抽出元文書へのリンクの3つ組、または項目名・他の用語との関係（同義語・反義語・例示、

Document Access Method via Specialized Glossary
Hajime Iwamoto†, Fumihito Nishino‡, and Toshihiro Kakimoto‡
†Kansai Electric Power Co., Inc.
‡Fujitsu Laboratories Ltd.

5W1H に基づくもの)・抽出元文書へのリンクの3つ組である。属性と関係については以下の種類のものを考えている。

[属性]
定義、訳語、性質
[関係]
(1)同義語、反義語
(2)例示、列挙
(3)いつ扱われたか（when）
＜時期＞＜期間＞等
(4)どのような活動の対象か（what）
<研究対象><開発対象><導入対象>等
<設置><改良><保守><点検><運転>等
(5)何の役に立つか（why）
<目的><効果>等
(6)関係する人物・組織（who）
<組織><部門>等
(7)関係する設備または場所（where）
<場所><設備>等

また、用語を分類整理するため、階層カテゴリの利用を検討している。
（例）[設備] － [発電所] － [発電機])
カテゴリは次のような効果を与えると考えられる。
① パターンのチェック、推測（適合率の改善）
② 用語集を見やすくする
③ ルールの増加防止

4.2 用語情報の抽出

用語の抽出のために筆者らが開発中の情報抽出エンジンを利用する[1]。この抽出エンジンは以下の特徴を持つ。
① 字面上のパターン分析により用語を抽出
② 辞書を用いない
③ 抽出パターン（抽出ルール）はヒューリスティックに得る

抽出ルールの集合は、既存の一般的なルール（略記等を抽出）に、抽出元文書群を分析して得られるルールを加えて構築される。また、用語の定義情報には正確性を期するため、一般の電子化用語集から抽出される定義情報を加えて補足する。

用語情報は XML 形式で記述し、テキスト処理の容易さとデータとしての汎用性を持たせる。

4.3 抽出実験

既存の一般ルールを用いて、関西電力社内の論文200 件から抽出した用語集をチェックした。また、抽出元文書について人手による分析も行った。
・抽出元文書群：水力変電部門の論文 200 件
・抽出用語情報：同義語（略記)、性質、活動対象（開発、研究、設計、製作、実用化）

結果を簡単に記す。
・用語数（無意味なものを人手で除く）：228 個
・用語情報数：246 個

用語集の分析及び抽出元文書の幾つかをチェックした結果、次の事項が確認できた。
・用語情報の一覧表示や検索により、用語のゆらぎや誤りが検出できる。
・２文書からの抽出用語情報の数が充分多く、その内容がほとんど同じ場合、抽出元の２文書の内容には大きな類似性がある。
・文書の緒論と結論からの抽出情報が多い。
・短い論文、専門家が執筆した論文からの抽出情報が少ない。
・用語の定義や列挙の記述は少なく、パターンとしても抽出しにくい。定義抽出の困難さについては[2]に詳しく記述されている。
・図表タイトルの抽出も有効である。

5.システム構成

システムの構成を図１に示す。利用者はテキストエディタを用いて用語抽出ルールを編集する。用語抽出エンジンは、用語抽出ルールに基づき抽出元文書群から用語（属性と関係）を抽出して、抽出元文書へのリンクと共に XML 文書の形で用語集に格納する。利用者は用語集アクセス部の提供する WWW ページで用語の検索条件を入力する。用語集アクセス部は検索条件に基づき、用語集を検索してその結果（属性・関係・抽出元リンク）を HTML 文書化して利用者に示す。

また、利用者が作成中の文書に対する用語チェックを実現するモジュールも後で追加する予定である。

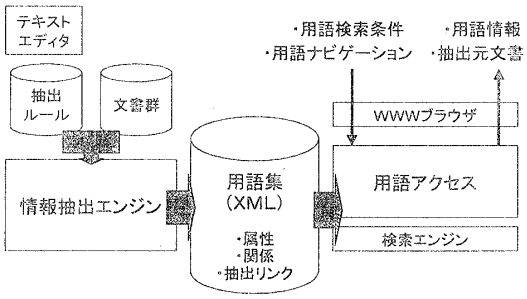

図１．システム構成

6.おわりに

今後、システムの設計を進め、実装・評価を行う予定である。

参考文献

[1] F. Nishino and et val, “Information Extraction using Top-down Pattern Analysis”, Proc. of 18th International Conference on Computer Processing of Oriental Languages, p.195-200, May 1999.
[2] 西野文人ほか, “テキストからの用語とその定義文の抽出”, 言語処理学会第５回年次大会, p.124-127, 1999