Japio

益々進歩する自然言語処理技術

国立国会図書館長
長尾　真

PROFILE
1994年に電子図書館アリアドネを公開。
2000～2007年、日本図書館協会会長。

自然言語処理技術の研究は１９６０年頃から始まったが、今日非常に多くの方法、アルゴリズムが研究され、それらを実現するソフトウェアが作られている。しかもそれらをパブリックドメインでだれもが使える環境が出来ることによって、より多くの人がそれらのソフトウェアの改良、新しいソフトウェアの作成に参加するようになって、ますますこの世界が豊富になってきている。

1 文献間の関連付け

特許申請文を審査したり、また特許内容を検討する場合に、類似の特許が既に存在するか、特許文章中の種々の記述に関係する学術文献にどのようなものがあるかを参照したいといったことがある。またぼう大な数の特許、特許申請文を取り扱う場合には、どのような分野についてのものかを特定して、その範囲において類似のものを探すといったことをしなければとても処理しきれないという状況がある。

したがってテキストの自動分類、類似のものをまとめるクラスタリング、検索を的確なものにするためのキーワード群の自動抽出、あるいは問題になる記述部分を発見するための全文検索といったことが必要となるが、これらについては種々の優れたソフトウェアが使えるようになって来ている。

しかしながら、技術が発展分化してゆくと分類１つをとっても数百程度の分類といった程度ではなく、トリー状にかなり深い階層をもったぼう大なカテゴリーの分類体系を構成することになる。したがってこのような分類体系の中に申請文を自動分類するためには、特許システムの分類体系に合った自動分類の技術を作る必要がある。また新しい技術分野が発生して分類体系に新しい項目を追加しなければならないといった場合にどうするかといった問題についてもさらなる研究が必要であろう。

類似の特許、関連する技術文献の検出とリンク付けについては種々のクラスタリング、類似性検出のための尺度が開発されて来ている。しかし特許文章に適した尺度についてはさらなる研究が必要であるうえに、ぼう大な数の科学技術文献データベースの整備と類似性検出については、その質やスピードなどに関して、さらなる工夫が必要である。

2 検索技術の高度化

グーグル検索が広く利用されるようになって検索技術についてあらためて注目が集まって来ている。グーグル検索の場合の問題点は相当注意深くキーワードを選んでも検索出力は何万、何十万と出て来ること、またその場合の出力順序が問題であることである。

特許関係文章の検索においては、検索用語を適当にコントロールすることが可能であろうが、関連する科学技術文献の検索までを含むとコントロールは難しくなる。また検索対象となるテキスト群はぼう大であるからグーグル検索と同じような問題に直面せざるをえない。

今日では検索ソフトウェアはいろいろと存在するが特

許文章に適した検索システムの開発が必要だろう。特に長い漢字複合語や漢字かなまじりの専門用語などがうまく扱えるシステムが大切である。

今日広く使われている各種検索システムの次の段階の検索システムは自然言語による質問文を受け付けるシステムであろう。検索語の集合によるマッチングでなく、複数の検索語の相互関係を導入した検索である。相互関係としては動詞を中心とした依存構造関係を用いることが多い。この依存構造関係でマッチングを取るのだが、これには相当な手間がかかる。したがってまず、質問文に現れる重要語をキーワードとして検索を行い、取り出されたテキストについて依存構造分析を行い、構造関係のマッチングを行う。こうすることによって検索の満足度は格段にあがる。これは現在京都大学の黒橋禎夫教授が開発中であり、少し強力な処理装置を必要とするが、実用にもってゆけるものである。

3 情報検索から事実検索へ

情報検索はキーワードを与えて、それに適合する本あるいは文献を取り出すシステムである。つまり情報の取り出しの単位が本や文献である。しかし目次検索や全文検索など種々の検索方式が開発されて来て、ある本のある章や節の数十ページ、あるいは該当するページやパラグラフ、あるいは文といったものを単位として取り出すことができるようになって来た。

そこで前節に述べた依存構造解析を伴なった自然言語文による質問のシステムを改良することによって質問文に対する答を直接取り出して回答するというシステムを作ることが出来る可能性がある。これを事実検索システムと呼んでいる。これまでの情報検索システムは質問者に対する答えを含んでいるであろう本や文献を提示するだけで、答えそのものは取り出した本や文献の中から探して下さいというものであったが、この新しい検索システムは質問に対する答えそのものを与えるという点において直接的である。これからの検索システムとして研究開発を強力に進めるべき課題である。

4 剽窃の検出

文字列マッチングのソフトウェアもいろいろと良いものが作られている。これは与えた文字列と同じものがテキスト中のどこに存在するかを検出するもので原則的にはテキストの大きさに比例した時間がかかる。

これに対してあるテキスト内の任意の長さの任意の文字列が他のテキストのどこかに存在するかどうかをチェックし、あればその場所を示すということが、これから必要となるだろう。誰かが他人のテキストのある文章部分を自分の文章の中にそのまま利用している、いわゆる剽窃を発見する時に必要となる技術である。

これは与えられた文字列がテキスト中のどこにあるかを探す文字列照合のアルゴリズムと比べると格段に難しい。またマッチングをするための手数は冪乗となり、ぼう大なスペースと時間を必要とするが、これを実用メモリサイズと実用時間内に行えるアルゴリズムを筆者らは十数年前に開発している。こういったアルゴリズムは工夫をすることによってさらに改良されたものが作れるだろう。また任意の長さの任意の文字列と類似の文字列が他のテキストのどこかに存在するかどうかを調べる、いわば近似的マッチングの技術を開発することも必要となるが、これはさらに難しい問題である。

特許文章だけでなく著作権関係などにおいても、この剽窃を発見することはこれからの大きな課題になるだろうから、この種の検索の研究を進める必要があると考えている。