平成 21 年 7 月 21 日

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

連想検索エンジン「GETAssoc」を公開
― 連想検索アプリケーションの構築を飛躍的に改善 ―

国立情報学研究所(所長：坂内 正夫)は、検索タスクに特化した新しい連想検索エンジン「GETAssoc」を公開しました。

「GETAssoc」は検索に頻繁に用いられる機能を中心に構成しているため、さまざまな用途に応じた連想検索アプリケーションを、従来に比べて極めて柔軟、かつ、容易に構築することができます。また、「GETAssoc」と連想検索アプリケーションの間は緩やかに結合するよう設計されているため、今までは、困難だったシステムやデータベースを統合することなしに、独立に運用されている異種の連想検索アプリケーションを連携させることにも成功しました。これによって、連想計算ライブラリの負荷低減を達成するとともに、「GETAssoc」のデータ保持モジュールの工夫により、従来難しかった対象データベースの文書の編集を容易にし、運用の柔軟性を大きく向上させました。

■GETAssoc 公開サイト:http://getassoc.cs.nii.ac.jp/

【連想計算とは】

連想計算は、文書に現れる単語の頻度に基づいて、類似文書や関連単語を統計的手法によって計算する技術です。連想検索は、連想計算を応用した文書検索アプリケーションで、ある文書から関連文書を探し出すといった高度な検索を行うことができます。さらに、文書が複数のデータベースに分かれて登録されている場合でも、あたかもそれらが同一データベースに登録されているかのように、文書同士の関連度で検索が行えます。そのため、キーワード串刺し検索などのように検索結果を統合する手法に比べて、より効率的に関連文書を探すことが可能になります。

このような優れた性質を持つことから、連想計算やその応用の連想検索は、企業内の文書検索だけではなく、「WebcatPlus」「想-IMAGINE」「BOOK TOWN じんぼう」などをはじめとする、多くの Web サービスで利用されています。

【GETAssoc の特長】

■複数データベース間の横断検索が可能

「GETAssoc」の特長の一つは、分散して配置された連想計算ライブラリ間の連携を可能にする通信規約(プロトコル)gss3 を備えていることです。gss3 は、複数の計算機に分散して配置されているデータベースを連携させ、これらのデータベース間のやり取りを行います。

gss3 は GETAssoc 用に新しく設計されたプロトコルで、連想計算の特長の一つである複数のサー

バに分散して配置されたデータベースを横断的に検索できるという機能を効率的に利用できます。

■コンテンツの権利保護が可能

gss3 プロトコルにより複数データベースの連携を行なう際には、「GETAssoc」が通信を仲介することで、連想計算ライブラリ間での直接の通信はおこらないようになっています。この設計によって、コンテンツの権利保護が可能になりました。例えば、あるデータベースを他のデータベースと横断的に連想検索できるようにするために公開しなければならない情報は、データベース中の各文書を特徴づける単語リストだけであり、本文そのものは非公開のままでの運用が可能です。また、連想計算ライブラリ同士が直接通信する必要がないため、アクセス制御などの連想計算ライブラリ運用の負荷が低減されます。gss3 はこのような特長を備えているため、gss3 に対応した連想検索アプリケーションであれば、既に gss3 を採用している他のアプリケーションなどとも連携させて運用することが容易です。

■連想検索アプリケーションの構築が容易

gss3 のもう一つの特長は、連想検索アプリケーションから連想計算ライブラリを効率的に呼び出すことができるよう設計されていることです。そのため、煩雑な呼び出し手順を記述することなく、容易に連想検索アプリケーションを構築することができます。例えば、これまでは、連想検索アプリケーションを実装する際、一回の連想検索のために複数回の連想計算ライブラリを呼び出す手続きを記述する必要がありましたが、gss3 プロトコルを利用すれば一回の呼び出しで必要とする計算を一気に行なえます。

【従来の連想計算ライブラリとの比較】

連想計算を高速に行なうライブラリには既に「GETA」があります。しかし「GETA」は、あらゆる応用を考慮した汎用の設計となっているため、可搬性は高いものの特定の応用を考えると取り扱いが容易とはいえませんでした。

「GETAssoc」は連想計算機能だけではなく、連想検索アプリケーションから連想計算を呼び出すための機能一式を備えていることが特長です。いわば、「GETAssoc」は、連想計算ライブラリと連想検索アプリケーションとのインタフェースと位置づけることができます。

＊「GETAssoc」は修正 BSD ライセンスを採用しており、商用などの利用も可能です。
＊配布は下記サイトよりダウンロードができます。
http://getassoc.cs.nii.ac.jp/

【付録】

その他の、「GETAssoc」の特長について。

1. gss3 の装備によって Web サイトなどへの組込みが容易に。

プロトコルの表現に XML を採用しており、きわめて容易に、汎用の XML パーザを備えたクライアントから「GETAssoc」を呼び出し、利用することができます。同様に、XML データの転送には HTTP を用いており、Java、Cocoa、perl、Ruby などの HTTP をサポートしているシステムであれば、煩雑なネットワークプログラミングを行なうことなく連想検索アプリケーションの開発を行なうことができます。また、JavaScript などを用いれば、Web ブラウザ上で動作する連想検索アプリケーションを簡単に作成することができます。

2. データベースのセットアップが容易に。

データベースを構築するために必要なファイルには、プレインテキスト、タブ区切りテキスト、XML など様々な形式が利用できます。特に、タブ区切りテキストの生成は、Excel などの表計算ソフトで作成した表形式のデータを指定された形式で保存するだけで生成できるため、コンピュータに関する専門的技術がなくとも容易にデータを準備することが可能です。

連想検索の性能を最大限に引き出すためには、連想計算に用いるべき対象データとして文書のどの部分を用いるのかを調整する必要があります。この調整は、「GETAssoc」では、連想計算の対象とすべき部分をテキストとして抜き出したファイルを作成するだけで十分です。テキスト形式であることの利点として、連想計算の対象部分の確認が容易なためきめ細かな調整が可能となり、その結果、連想検索の精度を最大限に高めることができます。

3. 全文を対象とした任意の部分文字列検索機能を持ち、その検索結果に限定した連想検索の実行(2 種類の検索を組合せること)が可能。

「GETAssoc」では、インデクシングされているキーワードに関する論理式を満たす文書や、任意の部分文字列を含む文書などに検索範囲を限定し、その範囲内の文書だけを対象にした連想検索を行なうことができます。

この機能を利用することで、検索サービスなどに絞り込み機能等を付加することが可能となります。

4. 既にセットアップしたデータベースに対して文書の追加・削除が可能に。

文書を随時追加できることで、対象データベースの文書を編集し、その編集結果を直ちに連想検索アプリケーション上で確認する、といった対話的な利用形態が可能になります。

連想計算では、文書間の類似度の他に単語間の類似度を計算する必要があり、そのためのデータを予め計算して保持しています。このデータの計算には手間がかかるため、文書の追加は難しいものでした。実際、「GETAssoc」と同様の連想計算が行える「GETA」は、連想計算の性能を重視しており、文書の追加はできないという設計になっています。「GETAssoc」はデータを保持しているモ

ジュールの工夫により、現実的な時間内での文書の追加機能を実現しました。また、それに伴う連想計算の性能低下はごく僅かとなっています。

5. 形態素解析器の切り替え・カスタマイズ機構をもち、ユーザによる形態素解析器のチューニング等が容易に。

連想検索を行うための最小単位は単語ですが、日本語では形態素解析器を用いて単語に分割する作業が必要です。連想検索で用いる形態素解析器のチューニングは性能向上のためにはきわめて重要ですが、これまでは専門的な知識を必要とするものでした。「GETAssoc」に付属する stmd というプログラムを利用することで、既存の形態素解析器を容易に組み込むことができるとともに、perl などの簡易プログラミング言語で形態素解析器の入出力の細かな調整を行なうことができます。

<<本件に関する問い合わせ先>>
国立情報学研究所 コンテンツ科学研究系 教授
西岡 真吾
連想情報学研究開発センター
E-mail：getassoc@cs.nii.ac.jp

<<報道に関する問い合わせ先>>
〒101-8430 千代田区一ツ橋 2-1-2
国立情報学研究所 広報普及チーム（担当：佐久間）
TEL：03-4212-2131 E-mail：kouhou@nii.ac.jp