●●● 技術レポート ●●●●●●●●●●

コンテクストサーチエンジン：ポストGoogle時代の検索エンジンをめざして


生産品質強化本部
技術研究センター長
稲垣陽一


生産品質強化本部
技術研究センター
瀬戸口光宏


コンサルティング
ビジネスユニット
ESTコンサルティング
センター
中村隆宏

1．はじめに

コンテクストサーチエンジン（Context Search Engine、以下CTX）は、Webページや学術論文など、大規模な文献集合を対象とする連想検索システムである。従来の全文検索エンジンとは異なるWHAT型検索機能を提供する。この検索エンジンは、株式会社小学館（以下小学館）とCACとの共同研究プロジェクトから生まれた。本論稿は、コンテクストサーチエンジンのコンセプト、実装技術、および実験結果について報告する。

2．検索エンジンの技術背景

著名な検索サイトGoogleは、2004年4月時点で、全世界43億弱のWebページを検索対象としている。また平成15年版の情報通信白書資料編（参考文献2）によると、JPドメインのHTMLファイル数は平成14年末で7,438万と推定されている。

Web文書集合の巨大化と歩みをあわせて、それを検索する仕組みも、人手による収集分類システムから自動化された大規模検索エンジンへと発展している。この発展は、主としてクローラと並列全文検索技術に負っている。クローラ（Crawler）とはインターネットを巡回して大量のWeb文書を自動収集するプログラムである。

並列全文検索技術は、複数のサーバーマシンを用いて検索要求を並列処理する技術である。クローラおよび並列化技術により、検索対象文書の大規模化と大量検索要求の処理が可能となった。

一方、検索対象となるWeb文書集合が巨大化した結果、検索キーワードを含む文書は、通常膨大な数になる。そこで現在主流の検索エンジンは、ヒットした文書のランキング方式に工夫をこらしている。

例えば、Googleが採用しているPageRankは「多数のページから参照されるページは良い。また良いページが参照するページも良い」という原則に基づいたスコアリング方式を採用している。ページの信頼度に応じて整列されるため、検索結果の先頭部分に適合する文書を発見できる可能性が高いとされる。

検索精度の指標として、適合率（precision）と再現率（recall）がある。適合率は検索結果中の正答文書の割合を示す。一方、再現率は全正答文書中で検索結果としてユーザーに返されたものの割合を示す。ユーザーに検索結果として提示する文書数を多くすれば、再現率を容易に向上させることができる。しかしインターネットのような膨大な文書集合を対象とする場合、長大な検索結果リストを返されても、ユーザーは困惑するだけである。

そこで再現率よりも適合率が重要な指標となる。PageRankに代表されるランキング方式は、インターネット検索の適合率向上に成果をあげた。検索エンジンの技術的発展の詳細は、参考文献3を参照されたい。

3．CTXのコンセプト

3.1 課題

検索エンジン発展の次のステップとして、我々はWHAT型検索に注目した。ここでWHAT型検索とは:

Q. 明治維新に関わった歴史上の人物は?
A. 坂本竜馬、西郷隆盛、…
Q. 抗菌作用が施された日用品にはどんなものがある?
A. 歯ブラシ、まくら、ふとん、…
Q. 金魚の名産地はどこ?
A. 熊本、大和郡山、…

といったQ&A形式の検索をいう。

インターネット検索を利用するユーザーが最終的に得たい情報は、検索要求に合致するページURLである。しかしその出発点には、WHAT型の問題意識が存在していることも多いと考えられる。大量のWeb文書を用いて、前述のようなQ&A形式の検索を実現することができるならば、従来の全文検索エンジンとは異なるツールを提供することが可能となる。

3.2 語共起の利用

WHAT型の検索を実現する鍵は、語の共起であると我々は考えた。語共起とは、複数の語が近接して同時に出現することを指す。言語研究分野では、従来、共起についての研究が行われ、語と語の間の共起の強さを測定する様々な方式が開発されている。

検索対象であるWeb文書集合に対して、キーワードが出現する箇所周辺を精査することにより、キーワードと強く共起する語を得ることができる。キーワードとの共起度をスコアとして語をランキングすることにより、WHAT型の検索を近似することができる。

3.3 カテゴリによる選別

ただしキーワードと共起するすべての周辺語を対象とするのは非効率である。検索意図とは無関係な語も多数、検索結果に含んでしまう。

WHAT型の質問文の多くは述部対象をカテゴリによって絞っている。

Q. 明治維新に関わった歴史上の人物は?
　キーワード: 明治維新
　共起カテゴリ: 歴史上の人名
Q. 抗菌作用が施された日用品にはどんなものがある?
　キーワード: 抗菌
　共起カテゴリ: 日用品
Q. 金魚の名産地はどこ?
　キーワード: 金魚
　共起カテゴリ: 地名

したがって対象となる周辺語を分類カテゴリによって選別することにより、ノイズを削減できる。

上記の考察のもと、語共起とカテゴリを検索基盤とするWHAT型連想検索エンジンCTXの研究開発を行った。

4．CTXの実装

4.1 システムの概要

CTXは、コーパスデータ作成部とコーパス検索部の2つのサブシステムから構成される。ここでコーパスとは、検索対象となる文献集合全体をさしている。

コーパスデータ作成部は、形態素解析された文書群を入力として、索引データと各語がID化されたデータを作成する。なお形態素解析には茶筅[*1]を用いている。

コーパス検索部は、キーワードとカテゴリ属性の組をユーザーから受けとり、キーワードと共起する語のうち、指定されたカテゴリ属性に属する語のみを、コーパスから抽出する。抽出された語はキーワードとの共起度によってランキングされ、検索結果となる。ユーザーは検索結果に対して、実際の出現文脈を確認することができる。

4.2 オンメモリコーパス

柔軟な検索要求に対応するため、オンメモリコーパスを実装アプローチとした。ここでオンメモリコーパスとは、すべての文献ファイルをコード化し、主記憶上で共起検索する方式をいう。

あらかじめ共起表を準備しておく方法は、3語以上の共起関係を求めたい場合に破綻する。例えば「花の写真に適したデジカメは?」というWHAT型検索は、2つのキーワード「花」と「写真」、および共起カテゴリ「デジタルカメラ」の3つの概念の共起関係を調べ、機種名のリストを返さなければならない。

任意の語、任意の語数の組合せについて共起関係をすべて事前に求めることは、現実的ではない。そこで、共起度の計算と共起出現箇所の抽出を主記憶上で高速に実行するために、文献ファイル集合全体を1つの配列として主記憶にマップした。低速なディスクストレージへのアクセスは、ロード時に限定する。

4.3 ランキング方式

抽出された共起語のランキング方式としては、単純頻度／tスコア／MIスコア／LogLogスコアの4種を実装した。それぞれ言語研究分野では著名な共起度の算出方式である。tスコアは、t検定の手法を応用して2つの語の共起強度を計る指標の1つ。コーパスの総語数をN_{corpus}とし、キーワードXと周辺語Yのコーパスにおける頻度をそれぞれN_XとN_Yとする。XとYの共起頻度をN_{XY}とすると、tスコアの算出式は:

$$t\text{スコア}=\frac{N_{XY}-\frac{N_X * N_Y}{N_{corpus}}}{\sqrt{N_{XY}}}$$

MIスコアは相互情報量ともよばれる代表的な共起度の算出方式である。キーワードと共起語それぞれのコーパス

*1) 奈良先端科学技術大学院大学 情報科学研究科 自然言語処理学講座が配布している形態素解析器である茶筅および共に公開されている日本語辞書システム。http://chasen.aist-nara.ac.jp/

における頻度によって共起頻度を割ることで、特徴的にノードワードと結び付く語が上位にランクされ、コーパスに多数回出現する高頻度語は逆に下位にランクされる。算出式は:

$$MI\ スコア = log_2 \frac{N_{XY} * N_{corpus}}{N_X * N_Y}$$

LogLogスコア（参考文献5）は、MIスコアに共起頻度の対数を乗じたもの。共起頻度をより積極的に評価する算出方式であり、頻度のみを考える単純頻度と特徴的な語を上位におくMIスコアの中間の尺度を与える。算出式は:

$$LogLog スコア = MI スコア * log_2 N_{XY}$$

後続の「5.実証実験と評価」の節で説明するが、一般的、常識的な概念を探したいときは単純頻度、「トリビアの泉」的な概念を見つけたいときは相互情報量を用いるなど、スコア方式によって意図に応じた検索が可能となる。

4.4　語の粒度

「語」の単位をどのように取るかは非常に重要な問題である。例えば、「東京大学」という語は、固有名詞として１語と考えられ、現実世界におけるエンティティに対応している。しかしカテゴリ辞書にこのエントリが存在しない場合、「東京」（地名）と「大学」（教育研究機関）の２語として扱われてしまう。

「自然言語処理において、有名な教育研究機関は?」という質問（キーワード=「自然言語処理」、共起カテゴリ=「教育研究機関」）に対して、「大学」という答えが返るのはナンセンスであろう。

すべての固有名詞や複合語がカテゴリ辞書に登録されているならば、上記の問題は発生しない。しかし現実にはそのような前提をおくことはできない。そこでCTXでは、連続する名詞自立語列を語の単位とした。

4.5　共起範囲

対象文書の量が膨大になればなるほど、ノイズを減少させるためには、検索範囲についての配慮が重要になる。例えば、語Xと語Yについてアンド条件でブーリアン検索を行うとしよう。長い文書の先頭に語Xが存在し、全く異なるトピックについて語っている末尾に語Yがある場合、その文書はヒットしてしまう。あるいは書籍のタイトルをリストしたページで、異なる行（すなわち異なる書籍）に含まれているXとYがヒットしてしまうケースも同様である。

CTXでは、語共起という局所的な性質を用いることにより、原理的にこのノイズを減少させることができる。さらに共起の対象範囲を形態素数/文数として指定することにより、ノイズを削減するためのさらに細かい調整を行うことができる。

4.6　事例によるカテゴリの指定

ユーザーはWebブラウザを介して、キーワード、カテゴリ情報、探索範囲、ランキング方式を入力する。指定したいカテゴリをカテゴリ辞書から直接選択してもよいが、カテゴリ数が数千というサイズになるとその選択は繁雑である。そこで、カテゴリの例を与えることで、カテゴリ指定の代替とするユーザーインタフェースを同時に実装した。

例えば、金属というカテゴリを「鉄」や「アルミ」などのカテゴリの事例となる語を入力することで代替することができる。事例によるカテゴリ指定インタフェースについては、5.2節でさらに説明する。

5．実証実験と評価

実証実験では、対象コーパスとしてWeb上の文書データを使用した。拡張子が「html」「htm」「txt」の文書データを対象として、文書数40万強（のべ形態素数３億強）を巡回回収し、システムの評価を行った。

ここでは、品詞情報を用いた検索と体系的な意味カテゴリ情報を用いた検索について、実証実験と評価報告を行う。

5.1　品詞情報を用いた検索

今回利用した形態素解析ツール「茶筅」の辞書IPADICでは、名詞が人名／組織／地域などの下位分類をもつ。この品詞情報を利用して、CTXの有効性を検証した。IPADICは固有名詞専用のカテゴリ辞書ではないため、本節の例はあくまで有効性検証のための試みと理解されたい。

「中田英寿と関係の深い人物について調べたい」という検索要求を例とする。キーワードを「中田英寿」、共起カテゴリを「人名」、共起範囲を形態素50（キーワードの前後それぞれ50形態素）としたところ、表１の結果が得られた。

単純頻度では、同じくサッカー選手である「中村俊輔」が最上位に現れるものの、スポーツ関連文書に多く現れる

表1　品詞情報を用いた検索の結果

順位	頻度数	LogLog
1	中村俊輔(18.0)	稲本潤一(54.3)
2	松井秀喜(15.0)	小野伸二(53.8)
3	小野（伸二）(14.0)	中村俊輔(52.4)
4	イチロー(12.0)	鈴木隆行(46.7)
5	小野伸二(12.0)	稲本（潤一）(44.2)
6	稲本潤一(12.0)	高原直泰(41.4)
7	稲本（潤一）(12.0)	高野進(37.9)
8	セ）リエ（A）(10.0)	西本聖(37.9)
9	鈴木隆行(9.0)	荒川静香(37.3)
10	ジーコ (8.0)	寺尾悟(37.3)

プロ野球選手（「松井秀喜」や「イチロー」など）も上位にランキングされる。一方、LogLogでランキングすると、サッカー選手が上位を占める。「中田英寿」との特徴的な結び付きの強さが影響していると考えられる。ちなみに「リエ」が表に現われるのは、「セリエA」を誤って形態素解析した結果、中間の「リエ」が人名として抽出されたためである。

以下に品詞情報を用いたその他の検索例とその結果を記す。

要求:「川崎市」と関係の深い組織は？

キーワード: 川崎市

共起カテゴリ: 名詞-固有名詞-組織

検索結果: 1.川崎(313.0)、2.富士通(170.0)、3.富士電機(148.0)［頻度数］

要求:「金魚」とゆかりのある地域を知りたい

キーワード: 金魚

共起カテゴリ: 名詞-固有名詞-地域

検索結果: 1.熊本(34.0)、2.日本(31.0)、3.大和郡山(14.0)［頻度数］

5.2 体系的な意味カテゴリ情報を用いた検索

前節では品詞情報に基づいて、3つのカテゴリ（人名/組織/地域）について検索する例をあげた。本節では、体系的な意味カテゴリ情報を用いた検索実験について報告する。

CTXに意味カテゴリ辞書をもたせることで、固有名詞以外の多様な種別の概念を対象とすることができる。

CTXの内蔵辞書として、階層状の意味カテゴリ辞書（オントロジ）を用意した。意味カテゴリ辞書は、例えば、「/生活・趣味/食品/食材-果物・木の実」といった階層構造をもつカテゴリと「芋」「トマト」「葡萄」などの語エントリから構成される。

ユーザーはリストからカテゴリを選択することができるが、カテゴリ数が1000を越える場合、その選択は必ずしも容易とはいえない。またユーザーにとって専門外の領域については適切なカテゴリを見つけること自体が難しいタスクと考えられる。そこで、カテゴリを直接指定するインタフェースの他に、検索対象カテゴリを直接指定せず、例を入力することで検索対象を指定するインタフェースを実装した（図1参照）。

このインタフェースを用いることで、例えば「リサイクルと関連の強い金属は?」という検索要求に対して、キーワードとして「リサイクル」、共起カテゴリの例として「鉄」を入力する。事例としての「鉄」は「/自然化学/化学/元素」「工業技術/建築/材料」などの共起カテゴリに変換される。

以下にこの検索例と結果を記す。

要求:「リサイクル」される金属にはどんなものが？

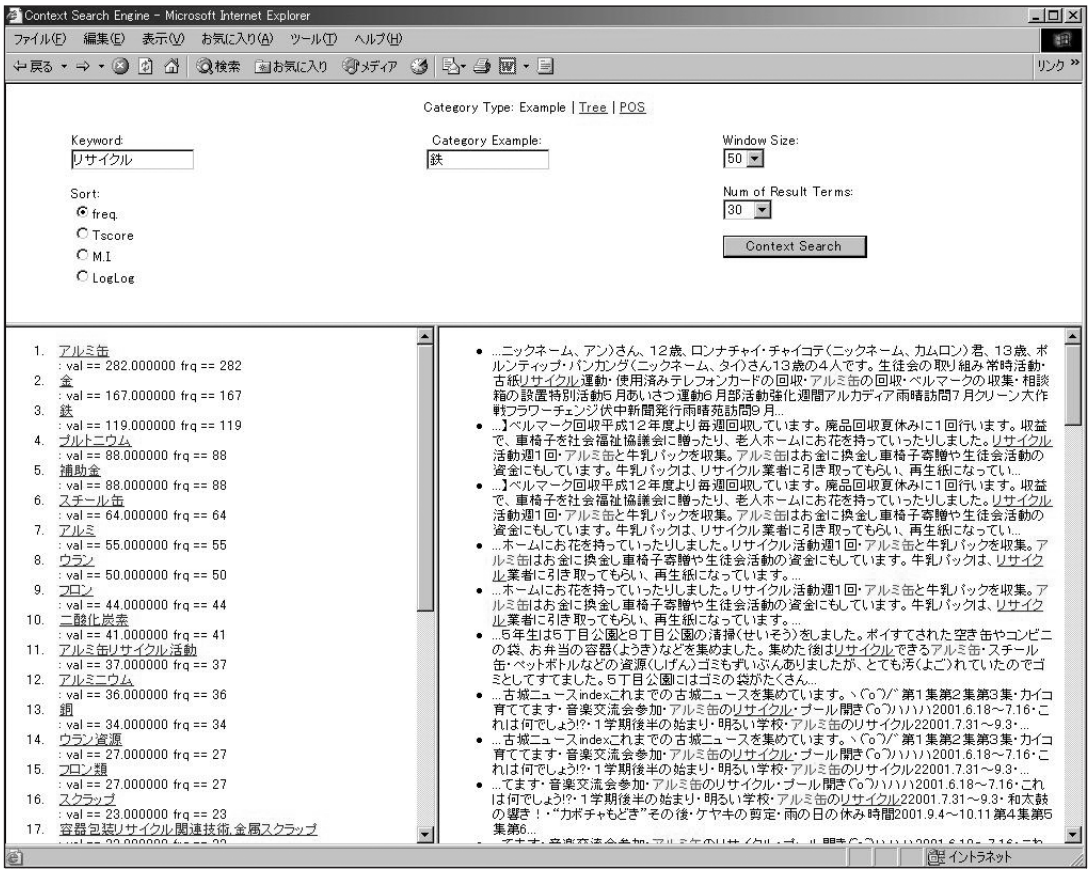

図1　カテゴリ例による検索

キーワード: リサイクル
共起カテゴリの例: 鉄
検索結果: 1.アルミ缶(282.0)、2.金(167.0)、3.鉄(119.0)[頻度数]
要求:「ポリフェノール」を含む果物と言えば？
キーワード: ポリフェノール
共起カテゴリの例: ブドウ
検索結果: 1.アメリカンチェリー(30.8)、2.カカオ(28.9)、3.リンゴ(24.3)
要求:「虫歯」と関係のあるお菓子は？
キーワード: 虫歯
共起カテゴリの例: ケーキ
検索結果: 1.ガム(20.0)、2.チョコレート(12.0)、3.まんじゅう(4.0)[頻度数]
要求:「癌」にかかりやすい体の部位は？
キーワード: 癌
共起カテゴリの例: 胃
検索結果: 1.大腸癌(265.0)、2.大腸(257.0)、3.肝(225.0)、4.食道癌(218.0)[頻度数]

5.3 考察

以上、共起とカテゴリを基盤とするWHAT型連想検索の検索例をみてきた。CTXの有効性を示すことができたと考える。

例を入力することで検索対象カテゴリを指定するインタフェースについても、ユーザー意図を十分に反映できると考える。また実証実験の対象文書数は40万ページと小規模であるとはいえ、エンジンは十分な実行性能を示している。並列実行形式にアーキテクチャを変更することにより、スケーラビリティも確保できるであろう。

文書データの収集についての課題（より多くの文書データの収集／専門特化した文書群を対象とした収集など）や、意味カテゴリ辞書、固有名詞辞書の整備に取り組んでいくことで、発見的な検索ニーズに応えることが可能となるであろう。

6．終わりに

以上、小学館とCACの共同研究から生まれたWHAT型連想検索エンジン「コンテクストサーチエンジン」について概略を報告した。現在、カテゴリ辞書の整備とエンジンコードのリファインを進めている。

CTX検索エンジンは、インターネット検索に限らず、多方面への応用が期待できる。

例えば、創薬系の論文データベースを検索対象として:
キーワード: 胃癌
共起カテゴリ: 遺伝子
とすると、文献集合において特定の病気と共起している遺伝子を関連度に応じてリストすることができる。新しい研究テーマの発見支援などに利用可能と考えられる。

今後、コンテクストサーチエンジンの応用領域を広く模索していきたい。

〈参考文献〉

1. 中島 睦晴、島田 博也：『インターネットコンテンツ統計に関する調査研究』郵政研究所月報（2002.9, pp.23-34.）
2. 総務省：『平成15年版 情報通信白書 資料編』情報通信統計データベース（http://www.johotsusintokei.soumu.go.jp/whitepaper/ja/h15/data/index.html）
3. 福島 俊一：『検索エンジンの仕組みと技術の発展』情報の科学と技術（54巻2号, 2004, pp. 66-71.）
4. 瀬戸口 光宏、稲垣 陽一、中村 隆宏、相澤 弘：『語の属性を用いた全文検索の高機能化』社団法人情報処理学会（第66回 全国大会（平成16年）、講演論文集(3)、pp. 55-56.）
5. A.Kilgariff and D.Tugwell：*WASP-Bench: an MT Lexicographer's Workstation Supporting State-of-the-art Lexical Disambiguation*, Proc. MT Summit XIII（2001, pp. 187-190.）
6. T.Nakamura and Y.Tono：*Lexical Profiling Using the Shogakukan Language Toolbox*, ASIALEX2003 Proceedings, The Third Asialex International Congress, August, Meikai Univ., Japan（2003, pp. 170-176.）