検索エンジンの仕組みを理解するための体験型教育用ソフトウェア

Software for Active Learning to Understand the Mechanism of Search Engine

松本このみ†，松浦敏雄†
Konomi MATSUMOTO and Toshio MATSUURA

概要　近年、中学校では教科「技術・家庭」、高等学校では教科「情報」として情報教育が必履修となった。当初、行われていた授業はコンピュータリテラシーに力点が置かれたものであったが、最近では情報の科学的理解についても重視されるようになってきた。一般に、情報の科学的理解を深めるための教材は、大学の授業資料を中心に様々な形で提供されているが、中・高校生向けの分かりやすい教材は十分とは言い難い。
本研究では、日常生活で重要な役割を果たし注目されている検索エンジンについて、中・高校生でも容易に理解できて、また実際に操作できる体験型のソフトウェアを開発した。本ソフトウェアでは、ブラックボックス化されているインデックスの作成過程と、単語を検索する際にそれがどのように活用されているのかを見て学ぶことができる。データの流れを目で追うだけではなく、任意のWebページをインデックス化できたり、利用者のペースでインデックス作成の過程を一つ一つ見ることができるなど、対話的に学ぶことができる。数人での模擬授業の結果、本ソフトウェアの有用性を確認した。

キーワード：検索エンジンの仕組み、体験型ソフトウェア、情報の科学的理解、デジタル教材
Keywords: Mechanism of Search Engine, Active Learning, Computer Science, Digital Materials

1　はじめに

初学者に情報の科学的理解を促すための教材として、ニュージーランドの Tim Bell 博士らが考案した CS（Computer Science）アンプラグドという手法が注目されている[1]。CS アンプラグドは、コンピュータを使わずに手や体を動かすことで情報の科学的理解を深めることができる教材であり、国内でも中学校を中心にいくつかの実験的授業が行われている。文献[1]では 12 の学習項目が用意されているが、もちろん情報科学の全てをカバーしているわけではない。特に検索エンジンの仕組みについては、日常生活で重要な役割を果たしていて注目されているにもかかわらず、CS アンプラグドには用意されていない。

検索エンジンの仕組みを学ぶ教材としては、大学での授業資料[2][3]や、Web ラーニングプラザ[4]のデジタル教材などが存在するが、中・高校生には内容が難しすぎる。中・高校生向けに映像を用いた分り易い教材として、NHK for School[5]があるが、自分で操作できるような教材ではない。情報の科学的理解を深めるためには、コンピュータ内部での処理内容流れを目で見ることができ、また実際に操作できるような教材が望ましい。

本研究では、検索エンジンの仕組みについて中高生でも容易に理解できて、また実際に操作できる体験型のソフトウェアを開発した。提案するソフトウェアでは、ブラックボックス化されているインデックスの作成過程と、単語を検索する際にそれがどのように活用されているのかを見て学ぶことができる。データの流れを目で追うだけではなく、任意の Web ページをインデックス化できたり、利用者のペースでインデックス作成の過程を一つ一つ見ることができる。

本論文ではまず、第 2 章で関連研究について紹介する。第 3 章では今回提案するソフトウェアの概要を説明し、第 4 章では、提案するソフトウェアを用いた授業の進め方について説明する。次に、第 5 章で提案するソフトウェアの評価を述べる。最後に、第 6 章で今回の研究のまとめと今後の課題について記述する。

† 大阪市立大学 大学院創造都市研究科

2　関連研究

2.1　CS アンプラグド

CS アンプラグドは、ニュージーランドの Tim Bell 博士らが考案したコンピュータ科学を教えるための手法である。コンピュータを使わずに体感を通してコンピュータの仕組みを学ぶことができるため、小学校や中学校を中心に授業が行われている[6]。文献[1]には全部で 12 の学習項目が掲載されている*1。CS アンプラグドのメリットは、教えるのが難しいと考えられているコンピュータ科学を、自分の手を動かしながらゲーム感覚で学ぶことができるため、小学生でも楽しんで取り組むことができるというところである。

大学コンソーシアム大阪などが主催の、中学生を対象とした 2 日間の CS アンプラグドセミナーでは、12 の学習項目のうちの 10 項目が実施された。セミナー中は、受講生は時間を追うごとに積極的になり、評価としても「コンピュータの仕組みの話をずっと聞いているだけではなく、自分の頭を使って考えながら取りくめたし、先生がわかりやすく説明をしてくれた」などの前向きな感想が多く、結果は満足のいくものであったという成果が示されている[8]。

2008 年度と 2009 年度には、情報オリンピック日本委員会と富士通が共同して、アンプラグドを利用した小学生向けのイベントが実施された。イベント終了後に子どもたちは、会場に併設されている博物館に移動して展示物を見学することで、学習したコンピュータの仕組みが実際の製品にどのように活かされているのかを積極的に確認していた。このことから CS アンプラグドの学習は、子どもたちにとって興味深い体験だったということが伺える[9]。

2.2　検索エンジンの仕組みを学ぶためのデジタル教材

検索エンジンの仕組みを学ぶためのデジタル教材として、大学の授業資料、NHK for School、Web ラーニングプラザについて紹介する。

2.2.1　大学での授業資料

授業資料[2]は、まず検索エンジンの仕組みをキーワードで紹介し、各仕組みの特徴については図のみで記載されているので、検索エンジンに必要な仕組みが一目で分かる。しかし、個々の説明文は記述されていない。授業資料[3]では、検索サイトの特徴や、検索エンジンの仕組みについて、それぞれの特徴を文章のみで簡潔に書かれているので分り易い。しかし、これらの資料は、いずれも中高生には内容が難しいといわざるを得ない。

2.2.2　Web ラーニングプラザ

小・中学生向けではないが、技術者向けに Web 上でコンピュータの仕組みの流れを学ぶことができる Web ラーニングプラザという Web ページがある。Web ラーニングプラザのメリットは、インターネット環境と一般的なプラグインソフトがあれば時間と場所を問わずに学習することができることである。 提供されている教材の分野は幅広く、各分野には複数のコースがあり、1 コースにつき 10 個程度のレッスンがある。また、1 つのレッスンにおいて、10～15 分程度のナレーションとアニメーションと、自己診断テストが用意されている。

Web ラーニングプラザが提供しているコースの中に、情報通信分野の情報検索コース[10]がある。このコースの学習の目的として「情報検索技術の理論的および実用的な背景の知識のみだけではなく、情報の要約や翻訳、質問応答といった関連技術も含めた、より広義の情報アクセス技術の体系的理解と、新たな情報システム構築や情報サービス提供のための基礎知識を獲得すること」が挙げられている。このレッスンでは、情報検索システムの構成を学ぶことができる。この学習において、情報を入力してから結果が返ってくるまでの一連のデータの流れをアニメーションでと音声で学習できる。

Web ラーニングプラザは、音声とアニメーションを利用したマルチメディア教材で比較的わかりやすいが、利用者が操作するようなインタラクティブな教材ではない。

2.2.3　NHK for School

NHK for School は、NHK が提供している学校

*1 日本翻訳も出版されている[7]。

向けコンテンツである。ここでは、1 つの教材につき 10 分間の動画が用意されている。このコンテンツの「ネットワークの活用」という教材では、検索エンジンについて学ぶことができる。

まず、「検索サービス」の説明にはじまり、Google が取り挙げられる。その後、検索結果の順位の決め方（ランキング方法）について、画像と音声で説明される。

NHK for School は、Web ラーニングプラザと同様に音声とアニメーションを利用したマルチメディア教材で比較的わかりやすいが、利用者が操作するようなインタラクティブな教材ではない。

3　提案ソフトウェアの概要

3.1　インターネットの情報

今回の研究では、初学者が既存の検索エンジンの仕組みについて、理解を深めることができるような教育用ソフトウェアの実現を目標としている。提案するソフトウェアは文献[11]を参考にしながら、主として Google の検索エンジンを中心に、その仕組みを解説している。しかし Google の内部処理は明らかにされていないので、Google の手法を忠実に再現しているわけではない。

検索エンジンには、前処理と後処理がある。まず、インターネット上にある膨大な量の Web ページをクローラというプログラムが収集し、リポジトリに格納する（図 1 参照）。その後インデックス作成を行う。

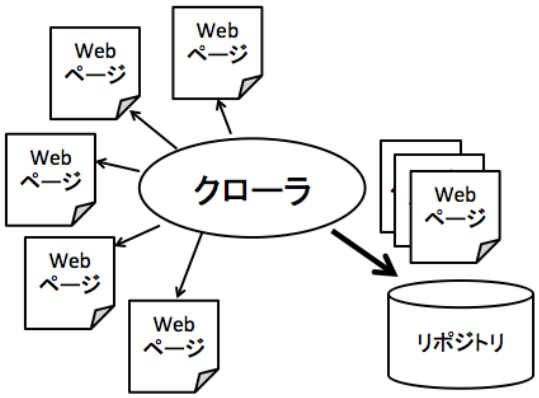

図 1 クローラの処理

インデックス作成の手順は 3 つある。まず 1 つ目は、Web ページ内の全ての単語に ID を割り振り、単語とその ID を一つにまとめた中間ファイルを作成する。この処理を Map と呼ぶ。次に、手順 1 でできた複数の中間ファイルをまとめ、単語の ID 順に並び変える。この処理を Shuffle と呼ぶ。その後、手順 2 でまとめたものを単語の ID ごとにまとめる。この処理を Reduce と呼ぶ。Web ページの収集からここまでを前処理と呼ぶ（図 2 参照）。その後、前処理で生成したものを利用して実際に検索を行う（図 3 参照）。これを、後処理と呼ぶ。

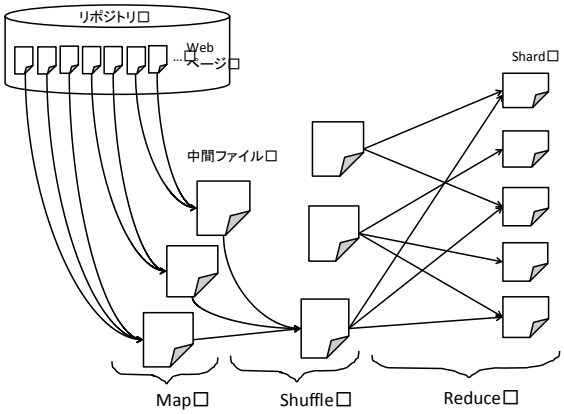

図 2 検索エンジンの前処理の流れ

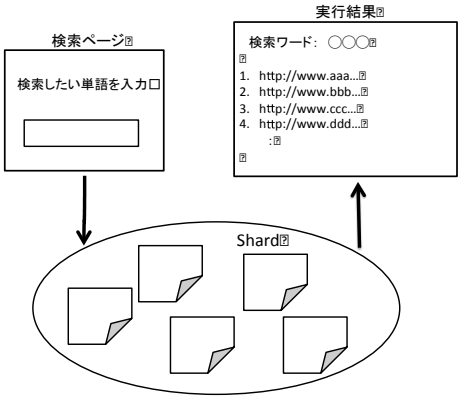

図 3 検索エンジンの後処理の流れ

なお通常では、クローラは自動で Web ページの収集を行うが、提案するソフトウェアでは指定した任意の Web ページを収集する。更に、収集した Web ページのインデックスを作成する際の Map から Shuffle、そして Reduce へのデータの流れや、後処理のデータの流れなど、実際はブラックボックス化している部分を画面上で確認できる。

ここまでを順に体験することによって、検索エ

ンジンの仕組みの理解を深めることができる。

4　授業の進め方

この章では、筆者らが想定している授業の進め方を説明する。まずソフトウェアを起動すると、タイトルと学習の目標が表示される。授業では学習の目標に目を通させた後、［START］を押して【検索エンジンの仕組み】に遷移する。

【検索エンジンの仕組み】は、検索エンジンの仕組みについて簡単な説明が記述されているページである。授業では、まず今から何をするのかということを大まかに理解させるため、説明文を読ませた後に、今回は生徒自身がクローラとなってWeb ページを集めてくることを説明する。そして、集めた Web ページに対してインデックス付けする処理の流れを順番に確認していくのだということを説明する。説明後は、［次へ］を押して【トップページ】に遷移する。

【トップページ】では、［Web ページの選択］と［Lexicon（用語集)］と［検索画面］の三つのボタンがある。授業では、Web ページの収集を行う前に、まず Lexicon にあらかじめ登録されている単語を生徒に確認させるため、［Lexicon（用語集)］を押して図 4 をポップアップ表示する。

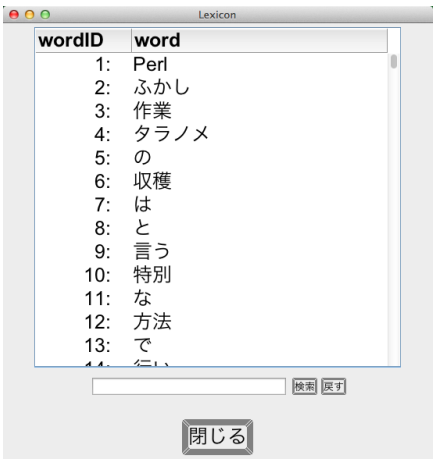

図 4 Lexicon

Lexicon は、wordID と word を保持している単語帳である。Lexicon では追加された単語順に wordID が割り振られるため、新しく単語を追加すると、Lexicon に登録されている単語の数 + 1 が新しい単語の wordID となる。

授業では、どんな word が格納されているのかを、スクロールバーを動かしたり検索ボックスに単語を入力させたりして生徒自身に確認させる。

ある程度の単語を確認させた後、Lexicon は後でも利用するので画面は閉じずに、【Map】の［Web ページの選択］を押して図 5 をポップアップ表示する。

図 5 Web ページの選択

図 5 では、任意の URL を追加できる。初期状態では 4 つのサンプル URL が入力されている。

授業では複数の URL を入力させた後、表示したいページの［閲覧］ボタンを押すと、選択した URL の HTML ページが表示される。

授業では、現在表示されているページを Map 処理するということを生徒に説明してから、この画面を閉じる。このページを閉じた後は図 5 に戻るので、次に［単語分割］を押して図 6 をポップアップ表示する。

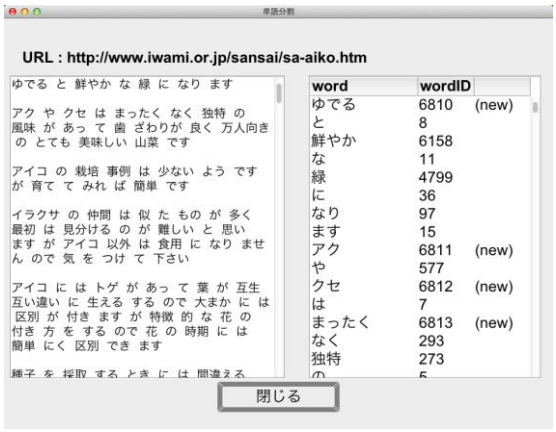

図 6 ページの分割

図 6 では、左側に図 5 で選択した URL の本文を分割したものが表示され、右側には、分割され

た単語とその単語の wordID が出てきた順番にリスト表示される。Lexicon に登録されておらず、該当する wordID がない場合は、新たに wordID を割り振ると共に、wordID 横に(new)と太字で表示される。

授業では生徒に、まず左側の単語と右側の表を見比べて、この単語にはこの wordID が割り振られているのだということを確認させてから、この画面を閉じる。その後は、図 5 に戻るので、次に［Map］ボタンを押して図 7 に遷移する。

図 7 Map の結果

図 7 では、図 5 で入力した URL を Map 処理した結果が表示される。各 URL にユニークな docID が割り振られ、それが URL 横に表示されている。

授業ではまず、［(docID)］（ここでは 216）を押して、図 8 をポップアップ表示する。

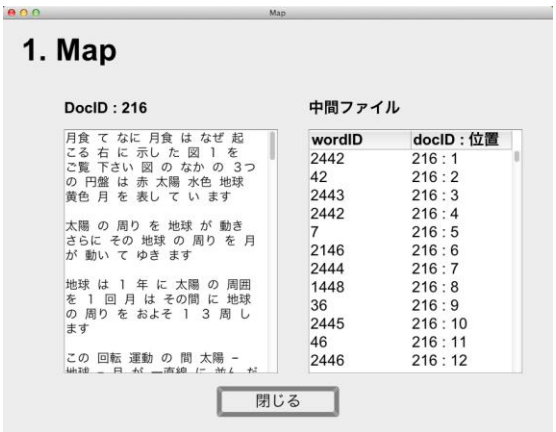

図 8 中間ファイル

図 8 では、図 7 で選択した docID の中間ファイルを見ることができる。今回は、前の図で 216 を選択しているので、docID ： 216 の中間ファイルが表示されている。左側には、docID の本文を分割したものが表示される。Map 処理は Web ページごとに単一で行われるため、複数のコンピュータがあれば、複数の Web ページを同時に Map することができる。

授業では、どのように中間ファイルが作られたのかを図 4 と一緒に見比べながら説明する。中間ファイルは wordID と、その単語が出現する docID：出現する位置(順番)という情報を保持しているので、中間ファイルを作成するには、最初に本文に出てくる単語の wordID を調べる必要がある。

まず左側の最初の単語を見て、その単語の wordID を Lexicon で調べる。そしてその wordID と、docID と出現する位置を中間ファイルの最初に格納する。ここでは、左側の最初の単語は「月食」で、Lexicon で調べると wordID は 2442 だということがわかる。docID は 216 で、「月食」という単語が出現する位置は Web ページの最初なので、中間ファイルには「2442　216：1」という情報が追加される。これを、5 つ程度順番に説明する。［閉じる］を押すと図 7 に戻るので、次に［次へ］を押して図 9 に遷移する。

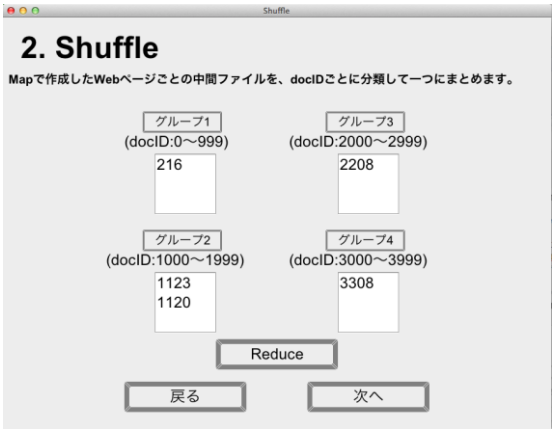

図 9 Shuffle

図 9 では、docID ごとにいくつかのグループに振り分けられる（上の図では 4 つ）。

授業では、それぞれの Shuffle の結果を確認させるため、各グループ名の右側にある［Shuffle］（ここではグループ 2）を押して図 10 に遷移する。

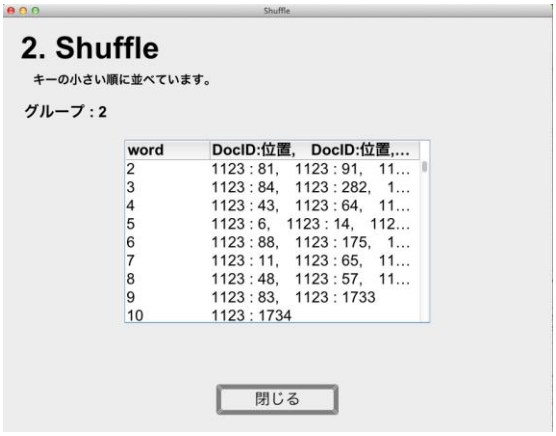

図 10 Shuffle の結果

図 10 では、グループ内の docID の中間ファイルを一つにまとめたものが表示される。今回は、図 9 でグループ 2 を選択したので、グループ 2 の Shuffle の結果が表示されている。

授業では Lexicon も利用して、ある単語はどの docID の、どの位置にあるかということを生徒に確認させる。例えば、wordID :4 の単語を Lexicon で調べると「タラノメ」だということが分かるので、「タラノメ」という単語は docID：1123 の 43 番目と、64 番目に存在するということが確認できる。ある程度確認させた後は、［閉じる］を押して図 9 に戻り、［次へ］を押して図 11 に遷移する。

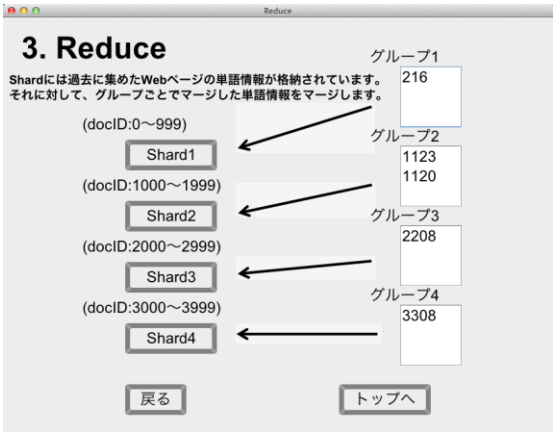

図 11 Reduce

図 11 では、現在 Google が実装していると考えられる Reduce の仕組みを示している。ここでは、Shard と呼ばれる各ストレージを用いて、docID の一定の範囲ごと（例えば 1000 ずつ）に分割した単語情報を保管している。この方法は、一つの単語情報が複数の Shard に分散して格納されているため、検索する際に並行して処理することができる。

授業では、それぞれの Shard の中身を確認させるため、［(Shard)］（ここでは Shard2）を押して図 12 に遷移する。

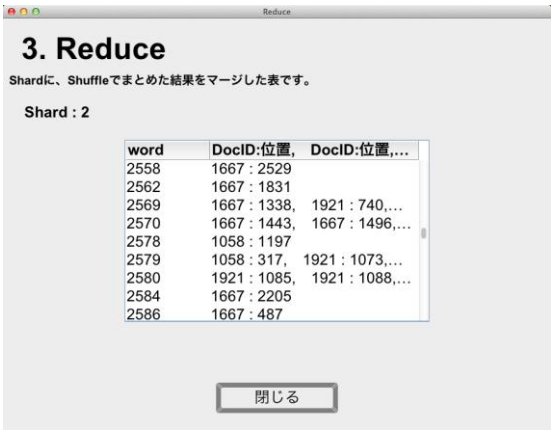

図 12 Shard の中身

図 12 では、Shuffle 処理でまとめた単語情報を追加した Shard の中身を確認することができる。今回は、図 11 で Shard2 を選択したので、Shard2 の中身が表示されている。この Shard が、インデックス処理の最終的な生成物となる。ある程度確認させた後は、［閉じる］を押して図 11 に戻り、［トップページへ］を押して【トップページ】に遷移する。そして、［検索画面へ］を押して【検索画面】に遷移する。

【検索画面】では、インデックス処理の生成物である Shard を利用して、実際に検索を行う。

授業では、Lexicon に登録されている単語を再度確認した後、その中から任意の単語をテキストボックスに入力させる（ここでは「浅草」と入力）。そして、［検索］を押して図 13 に遷移する。

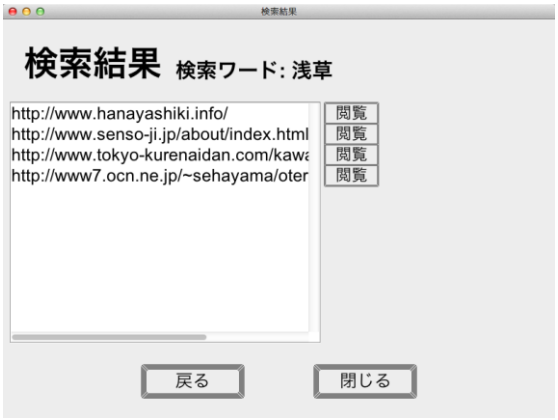

図 13 検索結果

図 13 では、【検索画面】で入力した単語の検索結果が表示される。今回は、【検索画面】で「浅草」と入力したので、その検索結果を表示している。

授業では、URL 横の［閲覧］を押して、選択した URL の HTML ページを確認させる。

5 評価

今回制作したソフトウェアの評価を得るために、大学院生 5 名を被験者として、模擬授業を実施した。授業は、4 章で述べた通りの進め方で実施した。被験者はいずれも情報科学の知識を有しているが、検索エンジンの仕組みについての知識は詳しくなかった。被験者の感想、意見を以下に示す。

今回の被験者は情報科学に関する知識を有しているので、生徒の立場よりもむしろソフトウェアの評価者としての意見を多くもらうことができた。

【感想及び意見】

・なんとなく仕組みが理解できた。
・説明が少なくて、今どの部分をやっているのか分からなかった。
・プログラム中、もしくは口頭での説明の量を増やせばもっと理解できそう。
・サンプル URL の他に、過去に入力された URL が確認できたり選択できれば、URL 選択の時に迷わなくてよいと思う。
・word と wordID が格納されている Lexicon の他に、URL と docID が格納されている DocIndex も閲覧できるようにすればいいと思う。
・Web ページ収集の際、この部分がクローラだと分かるように記載すればいいと思う。
・インデックスを作成した後の検索時に、単語を入力するとすぐに結果が返ってくるのは面白いが、作成したインデックスがどのように活かされているのかが見えない（検索時も、Lexicon や、Shard など、常に表示しておけばいいと思う）。
・検索エンジンに単語を入力してから結果が返ってくるまでのデータを流れが理解できた。

6 おわりに

今回の研究は、情報の科学的理解を促すための教材として、初学者が既存の検索エンジンの仕組みについて理解を深めることができるような、インタラクティブな教材を開発した。具体的には、通常自動で行われるWebページの収集を利用者自身が手動で収集する。更に、収集したWebページのインデックスを作成する際のデータの流れや、作成したインデックスを用いて検索する際のデータの流れなど、実際は隠されている部分を目で確認することができる。ここまでを順番に体験することによって、検索エンジンの仕組みについて理解を深めることができる。

実際にソフトウェアを利用してもらった結果、検索エンジンに単語を入力してから結果が返ってくるまでのデータを流れが理解できたなどの感想を得ることができた。また、wordとwordIDが格納されているLexiconの他に、URLとdocIDが格納されているDocIndexも閲覧できるようにすればいいと思うなどの具体的な改善提案を得ることができた。

今後の課題として、本ソフトウェアを中高生や文系の大学生に利用してもらい、多くの評価を得ることが挙げられる。

また、複数のキーワードでANDやOR検索ができるようにすることと、検索結果をランク付けして表示するための方法を組み込むことが挙げられる。

さらに、実際の検索エンジンでは多数の計算機が並列に動作して高速処理できていることが直接的に実感してもらえるように改良することなどが挙げられる。

□

参考資料

[1] Tim Bell, Ian H. Witten, Mike Fellows:

Computer Science Unplugged– An enrichment and extension programme for primary-aged children, 2006, http://csunplugged.com/~csunplug/books (2013/1/24 確認).
[2] 岡本 真: インターネットの特性(1) – 検索エンジンの仕組み, http://www.slideshare.net/arg_editor/otsuma2010427 (2013/1/24 確認).
[3] 正代 隆義: Google検索, http://colus.i.kyushu-u.ac.jp/~ts/04kiu/ouyou01.ppt (2013/1/26 確認).
[4] Webラーニングプラザ, 科学技術振興機構, http://weblearningplaza.jst.go.jp/ (2013/1/13 確認).
[5] NHK for School: 第3回 ネットワークの活用, 日本放送協会, http://cgi2.nhk.or.jp/school/movie/bangumi.cgi?das_id=D00051180043_00000&year=2012, (2013/1/24 確認).
[6] 兼宗 進, 久野 靖: コンピュータサイエンスアンプラグドの状況と今後の展開, 情報処理学会研究報告コンピュータと教育, Vol.2009-CE-98-23, pp.155-162, 2009.
[7] 兼宗 進ほか: コンピュータを使わない情報教育アンプラグドコンピュータサイエンス, イーテキスト研究所, 2007.
[8] 西田 知博: 中学生向けCSアンプラグドセミナーの実施とその課題の分析, 情報処理学会研究報告コンピュータと教育, Vol.2010-CE-106-3, pp.1-9, 2010.
[9] 西田 知博ほか: コンピュータ科学を楽しく学ぶ, 情報処理, 特集未来のコンピュータ好きを育てる, Vol.50, No.10, pp.980-985, 2009, http://kanemune.eplang.jp/pub/nishida090816.pdf (2013/1/13 確認).
[10] Webラーニングプラザ: 情報検索コース, http://weblearningplaza.jst.go.jp/のページより、[トップ] → [情報通信] → [情報検索コース] の順にたどる, (2013/1/21 確認).
[11] 西田 圭介: 『Googleを支える技術』技術評論社, 2008.
[12] Sergey Brin: Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, Vol.30, p.107-117, 1988, http://infolab.stanford.edu/~backrub/google.html (2013/1/19 確認).
[13] Luiz Andre Barroso, Jeffrey Dean, Urz Holzle: WEB SEARCH FOR A PLANET: THE GOOGLE CLUSTER ARCHITECTURE, Vol.23, p.22-28, 2003, http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/googlecluster-ieee.pdf (2013/1/19 確認).
[14] Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplied Data Processing on Large Clusters, p.137-150, 2004, http://www.cs.toronto.edu/~demke/2227S.12/Papers/mapreduce-osdi04.pdf, (2013/1/19 確認).
[15] 兼宗 進, 佐藤 義弘: 情報科教育法でのCSアンプラグドの利用, 情報処理学会研究報告コンピュータと教育, Vol.2010-CE-103-24, pp.1-3, 2010.
[16] 井戸坂 幸男, 兼宗 進, 久野 靖: 高校情報BにおけるCSアンプラグドの活用, 情報処理学会情報教育シンポジウムSSS2008, pp.201-206, 2008.
[17] 嘉田 勝: 情報科学の本質的理解を促す教育手法としてのコンピュータサイエンスアンプラグド, 教育処理学会研究報告コンピュータと教育, Vol.2010-CE-105-4, pp.1-5, 2010.
[18] 松本このみ: 検索エンジンの仕組みを学習するための体験型教育用ソフトウェア, 大阪市立大学大学院 創造都市研究科 都市情報学専攻 創造都市研究科 修了論文, 2013.