エンジニアリング・デザイン
＜学内検索エンジン＞

2011SE143 久野友菜
2011SE228 佐藤亜衣梨

1.要求定義

（担当：久野友菜、佐藤亜衣梨）

現状

現在南山大学の学内webページには検索エンジンが存在しておらず、リンクでたどり、ページを見つけるしか方法がない。探すのが面倒、またページの構成に慣れていない新入生にとっては情報を見つけ出すのが困難という問題が存在する。

解決策

・南山大学の学生を対象にした学内ページをキーワードで検索できる検索エンジンを作成する。
・検索エンジンをWebのブラウザで見れる形にする。
・情報が常に最新のものであるように、定期的に同期する。
・ひとつの単語のみでなくand検索、or検索などもできるようにする。
・学外からもページを確認できるようにする。

ターゲットの設定

ターゲットは南山大学瀬戸キャンパスを利用する学生に設定した。とくに2013SEや2013PPの1年生、またこれから入学してくる新入生の学生をターゲットとした。

対象ページ

・講義資料(Lecture Notes)
・情報理工学部（数理情報学部）・数理情報研究科
・S-AXIA利用の手引き(S-AXIA User’s Guide)
・瀬戸キャンパス用自動プロクシ設定ファイル(Automatic Prosy Configuration file)

2.設計

2.1.要求仕様書

（担当：久野友菜）

・学内のページをキーワードで検索できる。
・学外のページはヒットしない。
・利用者は検索ボックスにキーワードを入力する。
・「Search!」ボタンを押すとキーワードがシステムに送られる。
・システムはNamazuを利用してキーワードを全文検索する。
・システムはキーワードが含まれた学内ページを表示する。
・キーワードの部分は赤色で表示する。
・管理者は定期的にrsyncを利用してページの更新を行う。
・自分のパソコンにウェブサイトのデータが入りきらないため、入った分のみ検索にかけることができる。

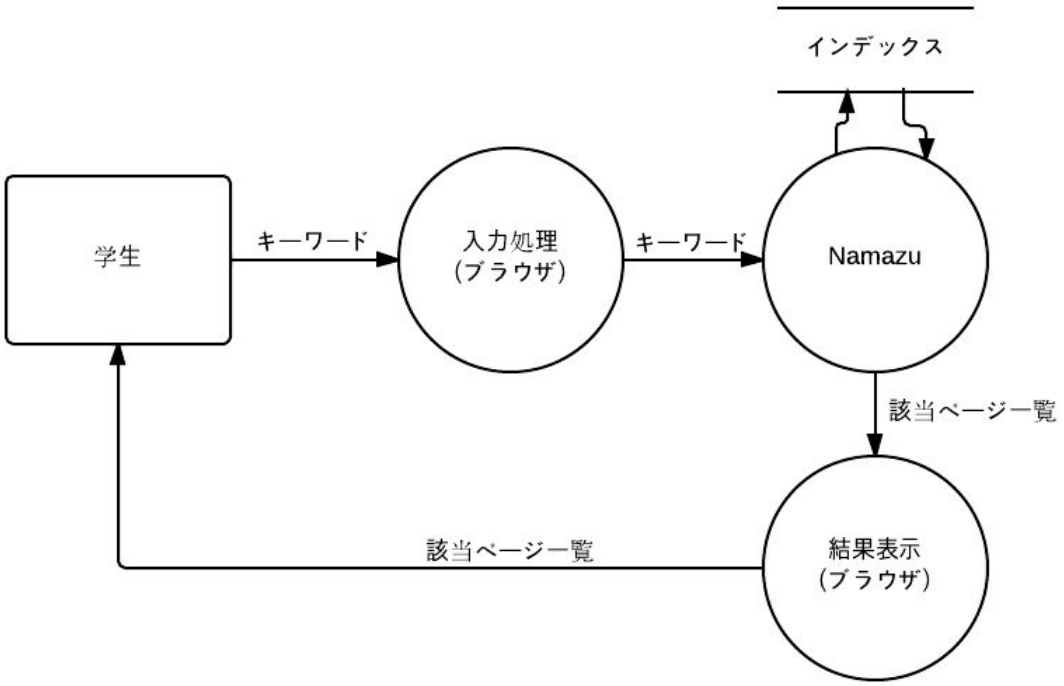
2.2.コンテキスト図

（担当：佐藤亜衣梨）

要求仕様書を元にコンテキスト図を作成する。

2.3.DFD

DFDを作るにあたり、検索エンジンシステムを「入力処理（ブラウザ）」「検索処理（Namazu）」「結果表示（ブラウザ）」の三つに分けた。
以下の手順でシステムを実行させたい。
(1)学生が検索ページのブラウザに検索したいキーワードを入力する
(2)ブラウザはそのキーワードをNamazuに送る
(3)Namazuはインデックスファイルの中からキーワードを検索する処理を行い、その結果をブラウザに出力する
(4)ブラウザは学生に結果を返す
以上より作成したDFDは以下のようになる。



3.実装

構成技術

動作環境：Linux
・rsync3.0.7
動作環境：Windows OS
・全文検索システム Namazu2.0.4
・漢字かな変換プログラム kakasi2.3.4
・ActivePerl v5.6.1
・Apache HTTP server 2.2

学内ページファイルの同期作業

（担当：久野友菜）

学内ページの同期にはrsyncを利用した。rsyncとは差分符号化を使ってデータ転送量を最小化し、遠隔地間のファイルやディレクトリを同期するアプリケーションソフトウェアである。10月頃はsshfsを利用しリモートサーバーをマウントする考えだったが、ファイル容量が足りない問題があり、断念した。現在は手動でrsyncで同期を行うことにしている。[1]
rsyncで学内ウェブサーバーから自分のコンピュータへウェブページデータを同期するコマンドは以下のようになる。

rsync -r 11se228@www-p.seto-private:/wwwfs/www-p/wwwdata/htdocs/classes/ise/2013

Namazu

（担当：佐藤亜衣梨）

rsyncで同期したページを検索可能にするために全文検索システムNamazuを利用することにした。Namazuの使い方はインターネットで検索した。[2][3][4][5]様々なページを参考にし、Namazuの実装を行った。当初はrsyncと同じLinuxで作業を行う予定だったが、Apacheの動作に問題が生じたためWindows OSに作業を移行した。Namazuの動作に必要なKAKASIとActivePerl[6]をダウンロードした。KAKASIは漢字仮名まじりの分をひらがなやローマ字綴りの分に変換するプログラムであり、単語ごとに分かち書きができる機能を利用する。
Namazuのメリットはあらかじめインデックスという、ファイル中の文字列と場所を調べておいたファイルを作成するため、ファイルの量が膨大な場合でも高速で検索処理を行える点にある。しかし検索対象の学内ページが更新されるたびにインデックスも更新しなければならないのがデメリットである。インデックス作成はコマンドプロントで行った。*c:\>mknmz -s -U -O c:\namazu\htdocs\index c:\namazu\htdocs\ED* というコマンドで*c:\namazu\htdocs\ED* の中にある学内ページのファイルのインデックスをc:\namazu\htdocs\indexの場所に作成した。検索はこのインデックスに対して処理を実行するため、コマンドで検索を行う際は（この場合の検索キーワード：store）*c:\>namazu store c:\namazu\htdocs\index* と入力する。
WebサービスにはApache[7]を利用し、http://localhost/cgi-bin/namazu.cgiに検索ブラウザを作成した。

検索ブラウザトップページ http://localhost/cgi-bin/namazu.cgi

4.テスト

4.1.コマンドでの検索に関するテスト

(1) 英語キーワード「linux」の検索を実行

c:\>namazu linux c:\namazu\htdocs\index
（実行結果は長いので割愛）
⇒20件の検索結果を表示（コマンドでは最大20件しか表示できない）

(2) 日本語キーワード「南山」の検索を実行
c:\>namazu 南山 c:\namazu\htdocs\index
検索結果

参考ヒット数: [南山: 2]

検索式にマッチする 2 個の文書が見つかりました。

1. 情報社会におけるソフトウェアの役割(蜂巣): 課題 (スコア: 2)
著者: hachisu@se.nanzan-u.ac.jp
日付: Tue, 07 Jan 2014 16:20:22
課題 問題1 問題2 問題3 問題4 自分の氏名をローマ字で記述し,ハフマン符合で表せ. また,ハフマン木とデータの圧縮率も示せ. ローマ字はすべて小文字を使い,空白文字は無視してよい. 解答例(必ずしも以下のように
/c|/namazu/htdocs/ED/2013/62253-001/hachisu/theme02S01.html (2,488 bytes)

2. 情報社会におけるソフトウェアの役割(蜂巣): 課題 (スコア: 2)
著者: hachisu@se.nanzan-u.ac.jp
日付: Tue, 07 Jan 2014 16:20:22
課題 問題1 問題2 データ生成のアルゴリズム 次の「データ生成のアルゴリズム」で作成したデータを選択ソートで降順(大きい順)に並べ替える手順を書け. 自分の氏名をローマ字の小文字で表す. ローマ字のアルファベ
/c|/namazu/htdocs/ED/2013/62253-001/hachisu/theme03S04.html (10,301 bytes)

現在のリスト: 1 - 2
⇒2件の検索結果を表示

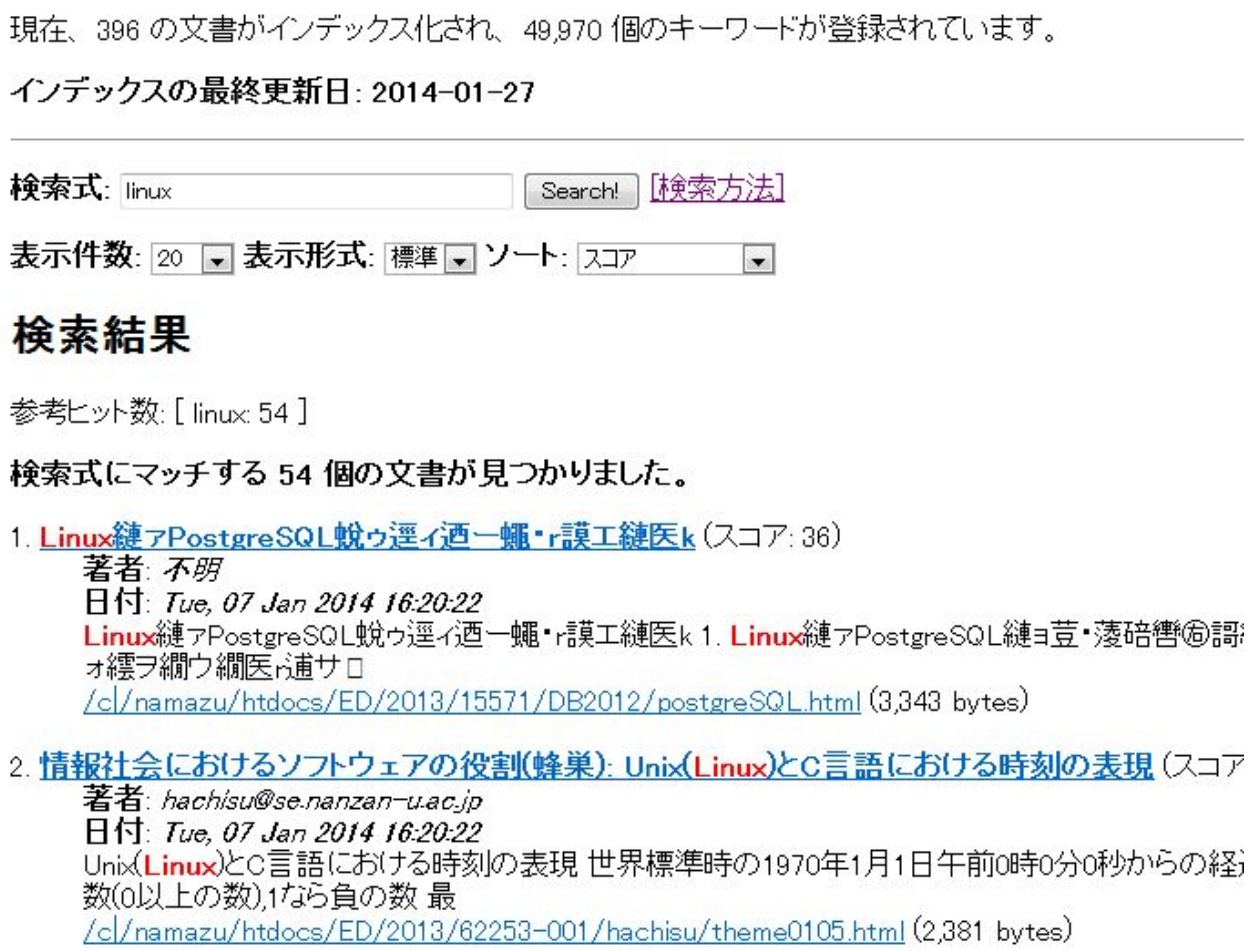
(3) 空欄の検索を実行
c:\>namazu c:\namazu\htdocs\index
検索結果

参考ヒット数: [c:\namazu\htdocs\index: 0]

検索式にマッチする文書はありませんでした。
⇒0件の検索結果を表示

4.2.ブラウザでのテスト

(1) 検索式に英語キーワード「linux」を入力し、Search!ボタンをクリック

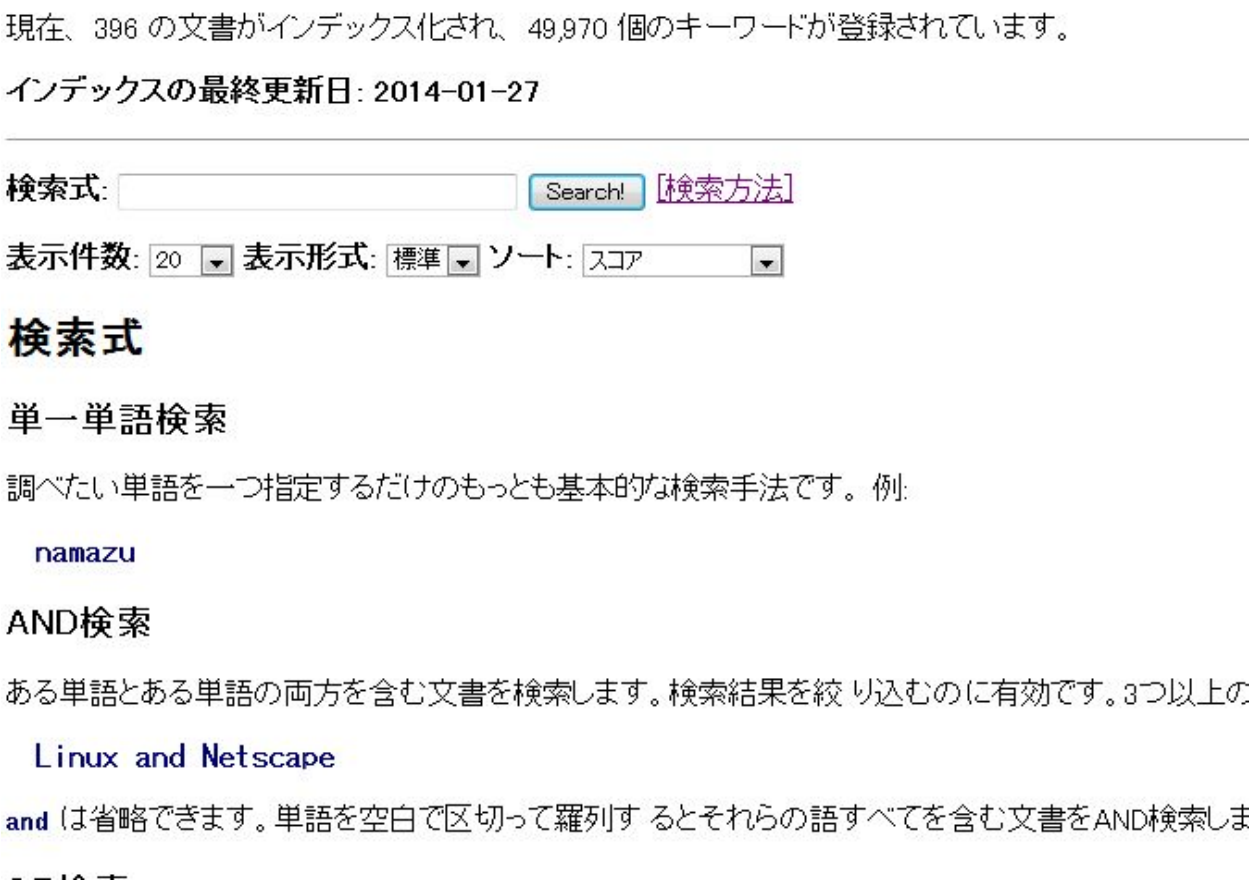


⇒54件の検索結果を表示

(2) 検索式に日本語キーワード「南山」を入力し、Search!ボタンをクリック

⇒2件の検索結果を表示

(3) 検索式に何も入力せずSearch!ボタンをクリック



⇒トップページを表示

5.考察

最初に述べた解決策を元に考察を行う。

達成できたもの：

・南山大学の学生を対象にした学内ページをキーワードで検索できる検索エンジンを作成する。

・検索エンジンをWebのブラウザで見れる形にする。

・ひとつの単語のみでなくand検索、or検索などもできるようにする。

課題が残るもの：

・情報が常に最新のものであるように、定期的に同期する。

現在は手動で同期を行うしか方法がなく、常に最新の情報という状態を保つのが難しい。また、linuxの環境下で同期を行っているため、Namazuを動かしているWindows OSに同期のたびにフォルダを移動させなければならない。さらにインデックスの作成も手動で行っているため、常に新しい情報を提供するためには改善すべき課題が多く残る。

・学外からもページを確認できるようにする。

現段階では検索ブラウザは外からはみることができない状態にあるので、外部からの接続はできない状況にある。

その他これから行えると思われる改善点：
・一部文字化けが生じるファイルの対処（htmlの作成されたコードによって日本語の変換ができないファイルが存在する問題）
・PDFなどのhtmlファイル以外のファイルが検索不可能になっている問題の解決
・検索ページの見た目の改善

当初の目的である検索ページの作成は達成できたが、実用化するにはまだまだ改善すべき点が多すぎると感じた。時間配分や作業の分担がきちんとできていなかったのが反省点だと思う。また、自分の知識を増やすことも必要だと感じた。

6.参照

[1] rsyncの使い方　http://webos-goodies.jp/archives/51213844.html
[2] 簡単なNamazu設置と2.0.21へのバージョンアップ法(Windows版)
http://biwa28.lolipop.jp/msearch152/drugstore/namazu_bun.htm
[3] Namazu for WINのインストール http://www10.plala.or.jp/miyazawa/namazu/namazu.html
[4] 全文検索システムNamazu(Windows用)の設定覚書 http://sakaguch.com/SetNamazu.html
[5] WindowsでNamazu http://www.alles.or.jp/~oga/namazu/index.htm
[6] ActivePerlのインストール方法 - Windows で perl を使おう！
http://pocketstudio.jp/win/activeperl/
[7] WindowsXPにApache2.2をインストールしよう！
http://www.webdlab.com/guide/apache/apache_1-1.php