Marcin Kapłon
Węzeł Warmińsko-Mazurski
OŻK-SB

Olsztyn, 15.09.2013 r.

(z *podziękowaniami dla Pana Sławomira Rodziewicza za wsparcie informatyczne)*

CRAWLER – NARZĘDZIE ZBIERANIA DANYCH DO EWALUACJI WYDARZEŃ KULTURALNYCH w ramach badania OŻK-SB „KALENDARZ KULTURY”

OPIS I DZIAŁANIE PROGRAMU

I. UWAGI WSTĘPNE

Crawler jest jednym z narzędzi, które MOGĄ BYĆ wykorzystane do przeprowadzenia badania „Kalendarz kultury”. Badanie opisane zostało szczegółowo przez Barbarę Fatygę w artykule: INSTRUKCJA DO BADANIA OŻK-SB „KALENDARZ KULTURY” na stronie Obserwatorium. Znaleźć tam można opis metod i celów tego badania. Po raz pierwszy[1] w ramach projektu OŻK–SB badanie przy użyciu *crawlera* zostało przeprowadzone latem 2013 roku przez Zespół Regionalnego Węzła warmińsko-mazurskiego na internetowych stronach regionalnych, powiatowych i gminnych, zgodnie z przywołaną wyżej „Instrukcją”.

II. DEFINICJA

Crawler (czyli inaczej „robot internetowy” lub „robot indeksujący”) jest dowolnie programowalnym narzędziem informatycznym – programem, służącym zbieraniu informacji o stronach internetowych, w tym monitorującym zmiany w zawartości tych stron. Programowalność *crawlera* sprawia, że istnieje możliwość dowolnego ustalenia liczby stron, częstotliwości ich odwiedzin przez

1 Badanie „Kalendarz kultury” *de facto* pierwszy raz zostało wykonane w 2012 roku, w ramach projektu „Dynamiczna diagnoza stanu kultury Warmii i Mazur”, dofinansowanego przez MKiDN oraz Urząd Marszałkowski Województwa Warmińsko-Mazurskiego, jednakże wtedy było to badanie pilotażowe, a dane były zbierane i kodowane inną metodą; por. raport B. Fatygi, M. Dudkiewicz, P. Tomanka „Kultura pod pochmurnym niebem.”

crawlera, słów kluczowych poszukiwanych przez program w treści stron oraz sposobu raportowania zdobytych w ten sposób informacji.

III. RODZAJE STRON INTERNETOWYCH

Podczas badania koncentrowano się na stronach, mogących zawierać zapowiedzi i relacje z wydarzeń kulturalnych. Do tabeli „*sites*" (patrz niżej) wprowadzono adresy 400 serwisów internetowych, zwyczajowo prezentujących bieżące informacje m.in. z życia kulturalnego. Przede wszystkim były to strony internetowych wydań regionalnych i lokalnych gazet, regionalne strony radia i telewizji, strony instytucji kultury o zasięgu regionalnym i lokalnym, takich jak gminne ośrodki kultury i biblioteki gminne, a także portale turystyczne. Do badania wykorzystane zostały również wszystkie oficjalne strony urzędów gmin i powiatów.

IV. BUDOWA

Do napisania programu użyto języka skryptowego PHP. Wszystkie dane zostały zagregowane w bazie danych MySQL.
Baza ta składa się z czterech tabel:

1. „*Sites*" czyli strony, a właściwie serwisy lub portale internetowe

Tabela zawiera adresy sprawdzanych przez *crawlera* serwisów internetowych w układzie powiatowym. Tabela tego rodzaju musi zostać przygotowana nad wyraz starannie przez badacza. Szczególnie ważne jest, by odpowiednio przypisać poszczególne adresy stron konkretnym powiatom, gdyż otrzymane dzięki takiemu układowi wyniki znacznie ułatwią późniejsze kodowanie, analizę i interpretację zebranego materiału. W badaniu, przeprowadzonym w województwie warmińsko-mazurskim na przełomie lipca i sierpnia 2013 roku tabela – zgodnie z tym, co wskazałem wyżej - zawierała 400 wpisów.
Przykład wpisu w tabeli: (szczycieński)www.wbp.olsztyn.pl/~gbpdzwierzuty/

2. „*Keywords*" – czyli słowa kluczowe

Tabela zawiera listę słów kluczowych, których robot szuka na wybranych stronach. Na potrzeby badania zidentyfikowano i wprowadzono do tabeli 50 słów kluczowych, których można było potencjalnie użyć do opisu wydarzeń kulturalnych.
Wykorzystano następujące słowa lub grupy słów:
artystyczny, artysta, biesiada, dożynki, festiwal, festival, festyn, impreza, impreza artystyczna, impreza rozrywkowa, inscenizacja, jarmark, kiermasz, koncert, literacki, malarstwo, muzyka, muzyczna, piknik , plener, poezja, potańcówka, premiera, premierowy , projekcja filmu, publiczność, raut, recital, rzeźba, seans filmowy, pokaz filmowy, spektakl teatralny, spektakl, spotkanie autorskie, święto, sceniczny, pokaz tańca, turniej tańca, turniej, uczestnicy, warsztaty

artystyczne, wernisaż, widowisko, widownia, wydarzenie artystyczne, wystawa, występ estradowy, zabawa, zespół artystyczny, zespół estradowy.

Rzecz jasna, wyrażenia te nie wyczerpują treści wydarzeń kulturalnych, jednakże uznane one zostały za wystarczająco wskaźnikowe do odszukania właściwych stron.

3. *„Pages"* - czyli strony
   Tabela zawiera listę adresów wszystkich stron i podstron w serwisach odwiedzonych przez *crawlera*. W badaniu w województwie warmińsko-mazurskim zebranych zostało N= 72.712 wpisów.

4. „*Pages_to_send*" – czyli strony raportowane.
   W tabeli tej zawarto adresy stron i podstron, na których znaleziono co najmniej jedno słowo kluczowe. Z efektów tych poszukiwań były generowane i codziennie wysłane raporty. W ciągu 30 dni badania na Warmii i Mazurach zebrano w tej tabeli N= 18.015 wpisów.

V. Działanie *crawlera*

Proces przeszukiwania stron i wysyłania codziennych raportów składał się z trzech etapów. Za każdy etap odpowiadał oddzielny skrypt.

1. W pierwszym etapie *crawler* miał za zadanie odwiedzić i pobierać strony główne wszystkich serwisów zawartych w tabeli „*sites*". Następnie identyfikował i zapisywał wszystkie odnośniki, a następnie porównywał je z odnośnikami zapisanymi w tabeli „*pages*". Jeżeli danego odnośnika jeszcze nie było w tabeli „pages" (bo pojawiła się w serwisie nowa podstrona z nową informacją), to był on do niej dodawany z odpowiednim statusem, określającym, że jest to nowy adres, który będzie trzeba odwiedzić. Serwisy były odwiedzane kilka razy dziennie w odstępach 2-3 godzin. Dzięki temu zadaniu *crawler* potrafił przeszukać wszystkie podstrony w zdefiniowanych serwisach i najdalej w ciągu kilku godzin od pojawienia się nowej informacji (nowej podstrony) dodawał jej adres do zbioru „*pages*" – adresów wymagających przeszukania pod kątem występowania słów kluczowych.

2. W drugim etapie *crawler* zajmował się nowymi stronami dodanymi do tabeli „*pages*". Odwiedzał każdą z nich i przeszukiwał je pod kątem obecności słów kluczowych (z tabeli „*keywords*"). Jeżeli na danej podstronie znalazło się poszukiwane słowo, to adres tej strony i znalezione frazy były zapisywane do tabeli „*pages_to_send*". Zapis taki dokonywany był z nadaniem odpowiedniego statusu, dzięki któremu system rozpoznawał, czy informacja o tej stronie była już dodana do raportu czy nie.

3. Trzeci etap to wysłanie codziennego raportu z listą nowych podstron z tabeli „*pages_to_send*”. Raporty zawierają następujące informacje:
   - tytuł podstrony (często tożsamy z tytułem informacji) jako link aktywny;
   - adres podstrony z informacją o danym wydarzeniu kulturalnym;
   - znalezione słowa kluczowe.

   Przykłady wpisów z raportu:

   Ornat z krwi pod fromborską katedrą – Frombork http://frombork.wm.pl/164948,Ornat-z-krwi-pod-fromborska-katedra.html

   Znalezione słowa kluczowe: spotkanie autorskie, premierowy, koncert, literacki, impreza

   Eko - Piknik w Kochanówce http://gminalidzbark.pnet.pl/pl/aktualnosci/64-2013/1178-eko-piknik-w-kochanowce.html

   Znalezione słowa kluczowe: piknik

   DYSKOTEKA W AMFITEATRZE - Braniewskie Centrum Kultury http://www.braniewskiecentrumkultury.pl/576-BEZPIECZNE-WAKACJE-Z-BCK-DYSKOTEKA.html

   Znalezione słowa kluczowe: uczestnicy, muzyczna, muzyka

Pierwszy raport, zawierający wszystkie wystąpienia słów kluczowych na zindeksowanych podstronach, mieścił się na blisko 100 stronach programu Word. Następne były znacznie mniejsze, zawierały bowiem jedynie nowe wpisy. Raport, z którego pochodzą powyższe przykłady, wygenerowany został 25 sierpnia 2013 roku i liczył 29 stron.

***

Zebrane w opisany sposób dane poddane zostały dalszym etapom procesu badawczego – kodowaniu, analizom i interpretacji. Narzędzie to będzie oczywiście doskonalone i używane w badaniu „Kalendarz kultury” w jego kolejnych edycjach.