Projekt i realizacja agenta do okresowego monitorowania treści wskazanych stron internetowych i raportowania zmian treści na tych stronach.

Projekt na przedmiot **ISI**

Michał Kaszlej

4 lutego 2014

1 Dziedzina zastosowań agenta

1.1 Motywacja

Przystępując do implementacji agenta, chciałem tak wyznaczyć dziedzinę jego działania aby powstała aplikacja była czymś więcej niż tylko jednorazowym projektem na zaliczenie.

Takie założenie skierowało moją uwagę na problem, który w ostatnim czasie stanął przed grupą znajomych socjologów zaangażowanych w tworzenie projektu Obserwatorium żywej kultury (nazywanego też dalej OŻK).

1.2 Obserwatorium żywej kultury

Obserwatorium żywej kultury jest szeroko zakrojonym projektem skoncentrowanym na zbieraniu i gromadzeniu wiedzy i informacji pomocnych w badaniach socjologicznych.

Jednym z głównych składników projektu jest tworzenie ogólnopolskiej mapy przejawów żywej kultury[1]:

> *Żywa kultura to wielowymiarowe środowisko (milieu) życia jednostek i grup społecznych oraz funkcjonowania instytucji społecznych, w którym zachodzą dynamiczne procesy, rozwijają się praktyki kulturowe, powstają mniej lub bardziej trwałe rezultaty (materialne i niematerialne wytwory) praktyk. Zarówno jednostki, grupy, instytucje, procesy, praktyki, jak i ich wytwory charakteryzują się zróżnicowanym, najczęściej wielowarstwowym i zmiennym nacechowaniem aksjologicznym oraz zróżnicowanymi, zmiennymi i wielowarstwowymi, najczęściej polisemicznymi, znaczeniami.*

W takim rozumieniu każdy sklep mięsny, czy basen jest przejawem żywej kultury i informacja o jego istnieniu powinna zostać uwzględniona w OŻK.

Wiąże się to z dużą akcją pozyskiwania wiedzy - w którą (częściowo odpłatnie) zaangażowani są studenci.

1.3 Przyjęte metody pozyskiwania danych do OŻK

Dane pozyskiwane są przez uczestników projektu poprzez przeszukiwanie internetu i wypełnianie odpowiednich pól w arkuszu kalkulacyjnym.

Poniżej znajduje się wyciąg z instrukcji dotyczącej pozyskiwania danych dla OŻK:

> *5) Danych poszukujemy w następujący sposób:*
> *Generalna uwaga: przeszukujemy strony sięgając najdalej do 10 pierwszych adresów;*
>
> - *punktem wyjścia są oficjalne strony GMIN, oficjalne strony MIAST i POWIATÓW oraz LINKI do nich opisujące kategorie infrastrukturalne wymienione w tesaurusie;*
> - *niektóre gminy i powiaty mają bardzo dobre strony w wikipedii;*
> - *warto sprawdzić informacje z BIP, ale to raczej dla porządku*
> - *trzeba uważać i sprawdzać – np. na stronie powiatu jest link do sal bankietowych w powiecie; wchodzimy: sale są podane tylko w miejscowościach, bez zaznaczenia jaka to gmina – trzeba oczywiście sprawdzić gminę!*

- *dla wielu prywatnych kategorii infrastrukturalnych może być pomocna PANORAMA FIRM i lub REGON*
- *warto przejrzeć też strony i portale tematyczne: informacji kulturalnych dla województwa i powiatów, chórtownia, regiopedia, kultura ludowa itp.*

6) UWAGA : cały sens tego badania polega na STARANNOŚCI.

Ponadto każdy uczestnik odpowiada wyłącznie za ograniczony i określony zbiór gmin. Poniżej zamieszczam wykaz gmin przydzielonych mojej znajomej:

Miasta	**Gminy miejske**	**Gminy miejsko-wiejskie**	**Gminy wiejskie**
Góra Kalwaria	Pionki	Iłża	Gózd
Konstancin-Jeziorna	Sierpc	Skaryszew	Jastrzębia
Piaseczno		Góra Kalwaria	Jedlińsk
Tarczyn		Konstancin-Jeziorna	Jedlnia-Letnisko
Pionki		Piaseczno	Kowala
Iłża		Tarczyn	Pionki
Skaryszew		Lesznowola	Przytyk
Prażmów			Wierzbica
			Wolanów
			Zakrzew
			Gozdowo
			Mochowo
			Rościszewo
			Sierpc
			Szczutowo
			Zawidz

1.4 Zbierane dane

Zbieranie danych polega na wypełnianiu oddzielnego akrusza dla każdej gminy. Każdy arkusz zawiera następujące kolumny:

1. Nazwa gminy lub miasta na prawach powiatu
2. Numer TERYT

   Każdej gminie w Polsce przypisany jest numer TERYT.
3. Indeks kategorii

   Przypisanie obiektu do jednej z 28 kategorii, według zamieszczonej na stronie OŻK tabeli kategorii[2].
4. Kategoria własności obiektu

   Możliwe wartości:

   - P = prywatna
   - S = samorządowa
   - PA = państwowa (budżetowa)
   - N = NGO's (organizacje pozarządowe)
   - I = inne (np. mieszany typ, współprowadzenie itd.)
5. Nazwa obiektu
6. Adres obiektu
7. Telefon/Email

8. Liczba miejsc/Zatrudnienie
9. Liczba funkcji obiektu

   Cytat z instrukcji wypełniania:

   > *Wpisujemy symbol 1 = instytucja lub organizacja ma tylko 1 funkcję podstawową (np.: jest biblioteką , nieważne, że odbywają się w niej różne zajęcia); symbol 2 = pod jedną nazwą i adresem realizowane są 2 lub więcej funkcji (np.: w Centrum Kultury jest Kino wymienione z nazwy, a nie napisano, że Centrum ma salę kinową);*

10. PRZYPADKI WĄTPLIWE

    Możliwe wartości:

    - I = brak przypisania do kategorii
    - FW = brak informacji na temat formy własności
    - BD = brak części danych
    - NA = wątpliwość dotycząca aktualności danych

1.5 Ocena przyjętego w OŻK sposobu pozyskiwania danych

Ogrom pracy stojący przed członkiem projektu OŻK jest niewyobrażalny. Stoi on przed zadaniem zindeksowania wszelkich przejawów kultury na ogromnym obszarze.

Jego zadaniem jest przeglądanie internetu w poszukiwaniu pływalni, kościołów, sklepów spożywczych itp i skrzętne wprowadzanie ich do arkusza.

Jest to praca zakrojona na tygodnie, jeśli nawet nie miesiące.

Wydaje mi się, że znacznie lepszy efekt uzyskać można poprzez zastosowanie prostego agenta do pozyskiwania danych. Zaoszczędzony w ten sposób czas można poświęcić na weryfikację poprawności wyników jego działania.

Aby odzyskać zapracowanych znajomych, zdecydowałem się więc na takie właśnie określenie dziedziny implementowanego agenta.

2 Źródła pozyskiwania danych

Realizację projektu rozpocząłem od zbadania możliwych źródeł pozyskiwania danych.

Założyłem, że rozwiązaniem optymalnym byłoby uzyskanie możliwie dużej liczby danych przy możliwie niewielkiej liczbie źródeł. Założenie takie wynika z faktu, że im więcej źródeł, tym większy stopień komplikacji agenta.

Poniżej prezentuję kolejno sprawdzane przezemnie źródła danych. W tej chwili wykorzystałem tylko jedno z nich, niemniej - mysląc przyszłościowo -dodanie dowolnego podanego niżej źródła stanowiłoby wartościowe rozszerzenie funkcjonalności agenta.

2.1 KRS

Na stronie Ministerstwa sprawiedliwości znajduje się wyszukiwarka numerów Krajowego Rejestru Sądowego.

Dzięki tej wyszukiwarce możliwe jest w łatwy sposób znalezienie wszystkich instytucji zarejestrownych w Krajowym Rejestrze Sądowym.

Dalsze badania wykazały jednak, że formularz jest odpowiednio zabezpieczony, co komplikuje przygotowanie agenta.

Dużym minusem jest też fakt, że w ten sposób ograniczę pozyskiwane dane, wyłącznie do tych istniejących w KRS, pomijając wiele obiektów żywej kultury.

2.2 BiP

Biuletyn Informacji Publicznej zawiera bazę podmiotów wykonujących zadania publiczne. W szczególności są to instytucje państwowe.

Na stronie znajduje się odnośnik, który umożliwia pobranie danych wszystkich podmiotów w pliku xml.

Po analizie pliku okazało się, że baza BiP jest stosunkowo niewielka i przeterminowana - niewystarczająca jako samodzielne źródło danych.

2.3 Google

Zaimplemetowanie w agencie mechanizmów pozwalających na wyszukiwanie żywej kultury poprzez google wydawało się świetnym pomysłem.

Po krótkich badaniach, ze względu na stopień niezbędnej elastyczności agenta uznałem tę opcję za zbyt skomplikowaną.

2.4 Serwis zumi.pl

W końcu, inspirowany wymienioną w instrukcji do badania "Panoramą Firm" - trafiłem na serwis który zawiera już wszystkie poszukiwane przez OŻK informacje. Co więcej informacje te są dostępne publicznie oraz zorganizowane w łatwy do obsłużenia sposób.

Zumi zawiera dane obejmujące niemal wszystkie kategorie[2] poszukiwane przez OŻK. W serwisie znajdziemy zarówno urzędy państwowe, parafie, sklepy, warsztaty, pokoje do wynajęcia - dokładnie to czego poszukją członkowie projektu OŻK.

Dodatkowo format odnośników przyjęty przez portal, znacznie ułatwia wdrożenie agenta. Listę firm dla konkretnego miasta otrzymujemy poprzez analizę odnośnika:

```
http://www.zumi.pl/lista-firm/<województwo>-<miasto>
```

Przykładowo dla Tarczyna:

```
http://www.zumi.pl/lista-firm/mazowieckie-Tarczyn
```

3 Realizacja projektu

3.1 Implementacja

Projekt został napisany w języku Java. Do parsowania HTML wykorzystywana jest biblioteka jsoup. Dane zapisywane są w bazie SQlite z wykorzystaniem sterownika JDBC-sqlite. Zapewnione zostały też mechanizmy pozwalające na łatwe generowanie plików Excel z danych zapisanych w bazach.

3.2 Podział na podproblemy

Arkusz, wypełniany ręcznie przez członków projektu OŻK składał się z 10 kolumn (patrz sekcja 1.4). Kolumny te powinny być wypełnione danymi, które znajdują się w różnych lokalizacjach w sieci.

Aby uniknąć zbędnych komplikacji, proces pozyskiwania danych podzielono na pięć etapów:

1. Uzyskanie numeru TERYT
2. Uzyskanie listy obiektów
3. Skompletowanie danych obiektu
4. Przypisanie obiektu do kategorii
5. Zapis danych do bazy

W kolejnych sekcjach opisane zostaną szczegółowo poszczególne etapy.

3.3 Uzyskanie numeru TERYT

Numer TERYT uzyskujemy jednorazowo dla każdej gminy na samym początku przetwarzania.

Baza numerów TERYT udostępniana jest przez Główny Urząd Statystyczny, numer uzyskujemy korzystając z poniższego URL:

```
http://www.stat.gov.pl/broker/access/performSearch.jspa?searchString="+<nazwaGminy>+
"&level=gmi&wojewodztwo=<TERYTwojewodztwa>&powiat=&gmina=&miejscowosc=&advanced=true
```

Przy czym `<TERYTwojewodztwa>` odczytujemy wcześniej w analogiczny sposób. Numer TERYT jest przechowywany i dodawany do każdego rekordu gminy której dotyczy.

3.4 Uzyskanie listy obiektów

Listę obiektów (bądź podstron zawierających obiekty) kompletujemy na początku przetwarzania, jednorazowo dla każdej gminy. Możemy ją otrzymać korzystając z URL:
`http://www.zumi.pl/lista-firm/<województwo>-<miasto>`

W przypadku większych gmin/miejscowości dostępny jest podział firm według ulic. W tym celu parsujemy zawartość obiektu HTML o klasie `streetsInCity`.

Dla małych gmin, może się zdarzyć, że podział na ulice nie będzie dostępny. W takim przypadku zumi oferuje nam podział według poularnych branż - wyszukując obiekty z okolicy. W tym celu badamy zawartość obiektu HTML o klasie `popularCategories`.
W przypadku gdy badamy branże - może się zdarzyć, że obiekty które przeglądamy pochodzą z sąsiednich gmin. Niestety w chwili obecnej nie ma mechanizmu weryfikacji takich zdarzeń.

3.5 Skompletowanie danych obiektu

W następnej kolejności rozpoczynamy przetwarzanie kolejnych obiektów.

W wyniku poprzedniego kroku (uzyskanie listy obiektów) możemy od razu odczytać nazwę, adres oraz telefon obiektu. Lista zawiera też odnośnik do podstrony opisującej dany obiekt. Przykładowo:

```
http://www.zumi.pl/2961559,Browar_Tarczyn_Sp._z_o.o.,Tarczyn,firma.html#aboutPage
```

Z parsowania podstrony `#aboutPage` otrzymujemy email, formę własności oraz ilość zatrudnienia. Odczytana forma własności, jest dodatkowo mapowana wewnętrznie, tak aby w miarę możliwości pasowała do zdefiniowanych kategorii:

- P = prywatna
- S = samorządowa
- PA = państwowa (budżetowa)
- N = NGO's (organizacje pozarządowe)
- I = inne (np. mieszany typ, współprowadzenie itd.)

Często zdarza się taka sytuacja, że na stronie `#aboutPage` brakuje emaila, formy własności lub ilości zatrudnienia. Dla dwóch ostatnich przypadków uzupełniamy pole uwagi odpowiednio wartością "FW"(forma własności) lub "BD"(brak danych).

3.6 Przypisanie obiektu do kategorii

Lista kategorii - a właściwie tagów nadanych przez zumi - znajduje się w podstronie `#homePage`. Przykładowo:

```
http://www.zumi.pl/2961559,Browar_Tarczyn_Sp._z_o.o.,Tarczyn,firma.html#homePage
```

Kategorie te po parsowaniu, porównujemy ze stanem bazy danych **Slowniki**. Baza ta zawiera wprowadzone wcześniej kategorie oraz przypisany im indeks z tabeli kategorii przyjętych w OŻK.[2]

Dokładniejszy opis klasyfikacji znajduje się w sekcji 4.

3.7 Zapis danych do bazy

Dla każdej gminy tworzona jest oddzielna tabela w bazie danych SQlite. Aby umożliwić wielokrotne przetwarzanie jednej gminy, bez nadpisywania i kolizji nazw - nazwa tabeli ma formę:

```
<nazwaGminy><czasOdPoczatkuEryUnixa>
```

Tabela zawiera pola zawarte w sekcji 1.4

4 Klasyfikacja obiektów

4.1 Tabela Slowniki

W bazie danych istnieje tabela Slowniki, zawierająca mapowanie kategorii zumi na kategorie OŻK. Tabela składa się z dwóch kluczowych kolumn: NazwyKategoriiZUMI oraz IndeksuKategoriiOZK. Indeksy kategorii z OZK dostępne są na stronie OZK[2].

W początkowym okresie używania agenta, wymgane jest wprowdzenie niewielkiej liczby kategorii oraz przypisanie im mapowania na kategorie OŻK. Następnie, baza powinna uaktualniać się sama. W testach udało się z sukcesem klasyfikować przy bazie zawierającej ponad sześć tysięcy kategorii.

Zdaję sobie sprawę z faktu, że przy wielości możliwości klasyfikacji takie rozwiązanie jest bardzo prymitywne. Na jego obronę mam tylko fakt, że częścią zadań członków projektu OŻK jest rozszrzanie tabeli kategorii na stronie OZK[2]. Zebrany w celu klasyfikacji materiał z pewnością ułatwi im zadanie.

4.2 Przetwarzanie listy kategorii

Zakładamy, że mamy odczytaną ze strony i sparsowaną listę kategorii.

Wykonaj dla każdej kategorii:

- Jeżeli taka kategoria nie występuje w bazie Slowniki- oznaczamy ją jako kategorię do dodania.
- Jeżeli występuje ale z indeksem = -1 - oznaczamy ją jako kategorię do odświeżenia.
- Jeżeli występuje i ma indeks > 0 - zliczamy wystąpienie tego indeksu.

Wyznacz najczęściej występujący indeks kategorii i następnie:

- Kategorie oznaczone jako do dodania - dodajemy z tym indeksem
- Kategorie oznaczone jako do odświeżenia - odświeżamy z tym indeksem.
- Zapisujemy najczęściej występujący indeks jako kategorię obiektu

W ten sposób następuje automatyczne ukatualnienie tabeli Słowniki, oraz obiektowi przypisana zostaje kategoria.

5 Bibliografia

[1] http://ozkultura.pl/node/111 - Barbara Fatyga, definicja "żywej kultury"
[2] Barbara Fatyga, Eliza Gryszko, Aleksandra Orkan-Łęcka, INDEKSY INFRASTRUKTURY INSTYTUCJONALNEJ ŻYWEJ KULTURY, 2013 r.