ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO

NR 733 STUDIA INFORMATICA NR 30 2012

JOANNA MATUSIAK[*]
Uniwersytet Szczeciński

EKSTRAKCJA I AGREGACJA ZAWARTOŚCI STRON INTERNETOWYCH NA PRZYKŁADZIE PORTALI PRACY

Wprowadzenie

W ostatnich latach ekstrakcja i agregacja danych (ang. *web extraction*, *web scraping*, *web grabbing*) zyskują ogromną popularność. Powstaje też coraz więcej oprogramowań informatycznych, które w pewnym stopniu automatyzują ten proces. Niektóre z nich wymagają od użytkownika umiejętności programowania, oferując tym samym więcej swobody w ich adaptacji jako podkomponentów większych podsystemów[1], inne zaś oferują przyjazny użytkownikowi interfejs graficzny. Mechanizm ich działania opiera się zazwyczaj na decyzjach użytkownika, makrach internetowych i rozpoznawaniu wzorców, a przy tym oferują one możliwość zapisywania i przechowywania plików w powszechnie używanych formatach wymiany danych, takich jak: XML, XLS, CSV lub bazie danych[2]. Należy jednak zaznaczyć, że w przypadku złożonej ekstrakcji lub potrzeby terminarzowania zadań ekstrakcji w pewnych odcinkach czasu, ze względu na agregację danych, na przykład w celu zebrania większej ilości nowych danych pojawiających sie na stronach internetowych lub w celach statystycznych, narzędzia te nie zawsze mogą zostać wykorzystane z sukcesem.

[*] matusiak.joanna@gmail.com.

[1] http://web-harvest.sourceforge.net/.

[2] http://www.sundewsoft.com; http://www.lixto.com; http://www.newprosoft.com.

Ekstrakcja danych jest obecnie stosowana z dużym powodzeniem, na przykład do pozyskiwania danych z wielu różnych aukcji internetowych i agregowania ich w jednej bazie danych i w jednym systemie celem przeszukiwania ofert w sposób „globalny". Ponadto wiele firm komercyjnych pozyskuje informacje o cenach produktów/usług, na przykład konkurencji, aby w porę zareagować zmianą cen własnych produktów i stać się sprzedawcą bardziej konkurencyjnym dla konsumenta.

Aby zaprezentować praktyczne przeprowadzenie procesu ekstrakcji danych, wykorzystano oprogramowanie firmy SundewSoft o nazwie **WebSundew**. Do analizy wybrano strony internetowe dotyczące rynku pracy i oferowanych na nich nowych stanowisk pracy[3]. Analiza rynku pracy i pracodawców jest jedną z dziedzin, w których ekstrakcja i agregacja danych nie zostały wykorzystane w celu gromadzenia statystyk i jest to obecnie obszar znajdujący się w sferze badań.

1. Struktura WebSundew

Oprogramowanie WebSundew zostało zaprojektowane jako kolekcja wtyczek zintegrowanego środowiska programistycznego Eclipse[4], oferującego programistom tworzenie własnych rozszerzeń dzięki wbudowanemu środowisku tworzenia wtyczek PDE (ang. *Plug-in Development Environment*)[5]. Środowisko PDE oferuje możliwość tworzenia własnych komponentów, takich jak: perspektywy, widoki, właściwości, eksplorację plików projektu itp. powszechnie znanym użytkownikom środowiska Eclipse.

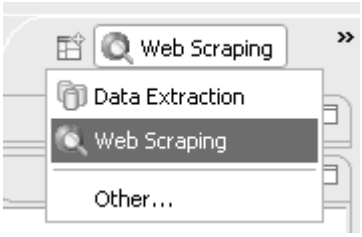
W celu ułatwienia pracy z programem wprowadzono koncept perspektyw. Przeznaczeniem perspektywy jest skompletowanie danej liczby zadań. Każda perspektywa posiada unikalną kolekcję widoków, które wyświetlają informacje na temat obiektów w sposób wygodny i czytelny dla użytkownika. Na przykład widok projektu wyświetla obiekty, które są dostępne w obecnie otwartym projekcie.

[3] http://www.pracuj.pl.

[4] http://www.eclipse.org.

[5] http://www.eclipse.org/PDE.

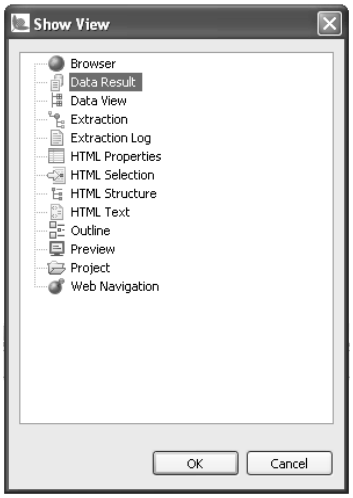
Lixto oferuje dwie następujące perspektywy:
- **web scraping**, która została zaprojektowana z myślą o wyszukiwaniu wzorców danych i wzorców następnych stron, jak również nagrywaniu makrointernetowych;
- **data extraction**, która jest przeznaczona do tworzenia źródeł danych, widoków danych i sekwencji ekstrakcji danych.



Rys. 1. Wybór perspektywy
Źródło: opracowanie własne.

Perspektywa web scraping zawiera następujące widoki:
- widok projektu (ang. *project view*);
- widok struktury HTML (ang. *HTML structure view*);
- widok właściwości HTML (ang. *HTML properties view*);
- widok selekcji HTML (ang. *HTML selection view*);
- widok tekstu HTML (ang. *HTML text view*).

Perspektywa data extraction zawiera:
- widok projektu (ang. *project view*);
- widok źródła danych (ang. *data source view*);
- widok wynikowych danych (ang. *data result view*);
- widok przeglądu danych (ang. *outline view*);
- widok danych (ang. *data view*).

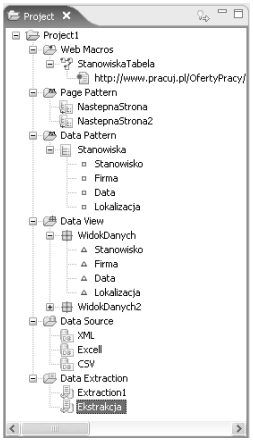


Rys. 2. Wybór widoku
Źródło: opracowanie własne.

2. Struktura projektu

Każdy projekt ekstrakcji i agregacji danych składa sie z następujących komponentów:

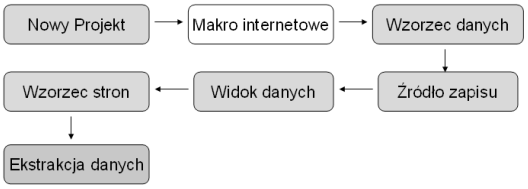
– makra internetowe (ang. *web macros*);
– wzorce danych (ang. *data pattern*);
– widoki danych (ang. *data view*);
– źródła danych (ang. *data source*);
– ekstrakcja (*a*ng. *extraction*).



Rys. 3. Struktura projektu ekstrakcji danych
Źródło: opracowanie własne.

3. Architektura przepływu informacji

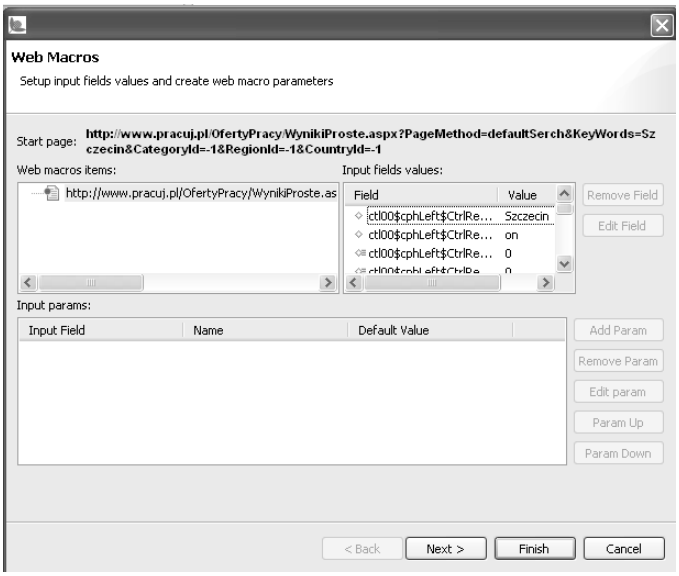
Jeśli uwzględnić dwie dostępne perspektywy, wiele dostępnych widoków i złożoną strukturę projektu, można stwierdzić, że przepływ informacji w oprogramowaniu WebSundew jest procesem dość złożonym. Na rysunku 4 przedstawiono w kolejności główne procesy składające się na sukcesywną ekstrakcję danych.



Rys. 4. Architektura przepływu informacji i wykonywanych procesów
Źródło: opracowanie własne.

W celu przeprowadzenia ekstrakcji i agregacji danych, należy:
- utworzyć nowy projekt;
- wskazać w **Web Navigator** adres strony internetowej;
- nagrać makrointernetowe;
- utworzyć wzorzec danych;
- utworzyć wzorzec relacji pomiędzy kolejnymi podstronami;
- utworzyć widok danych na podstawie wzorca danych;
- utworzyć źródło zapisywania uzyskanych danych z procesu ekstrakcji;
- utworzyć ekstrakcję, wybierając utworzone źródło zapisywania, wzorzec kolejnych podstron itp.;
- uruchomić utworzoną ekstrakcję danych.

4. Nagrywanie makrointernetowych

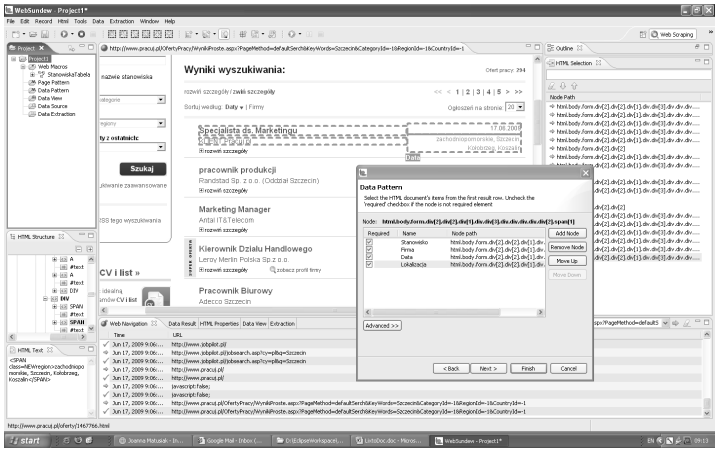


Rys. 5. Proces projektowania makrointernetowego
Źródło: opracowanie własne.

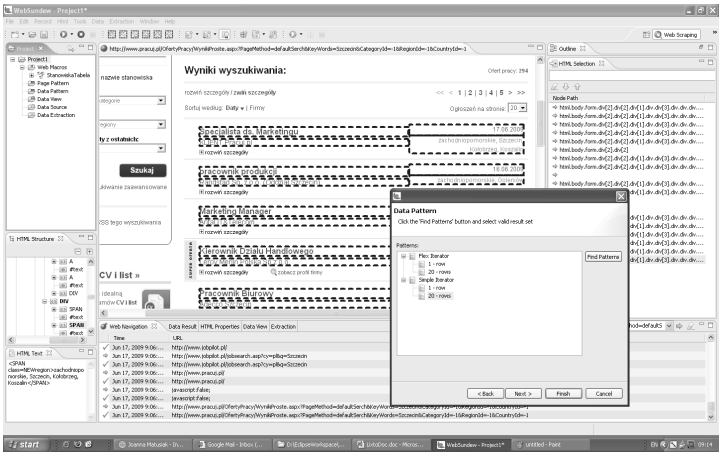
Pierwszym krokiem w kierunku ekstrakcji i agregacji danych jest nawigacja do strony, której dane staną się źródłem analiz. W tym celu wymagane jest makrointernetowe. Najprostszym przykładem makra jest nawigator do strony internetowej. Tego typu makra są używane, gdy dane są ulokowane na statycznych stronach. Adres internetowy strony (ang. *URL*) musi zostać podany w widoku nawigatora internetowego (ang. *WebNavigation*). W momencie, gdy strona internetowa zostanie załadowana, można rozpocząć proces nagrywania makra internetowego, klikając na **Play Web Macro** (rys. 5).

5. Wyszukiwanie wzorca danych

Wzorce danych wykorzystywane są do ekstrakcji danych ze stron internetowych. W celu wyszukiwania wzorca danych należy uaktywnić tryb selekcjonowania elementów strony przez selekcję kliknięciem myszy (rys. 6, 7 i 8).

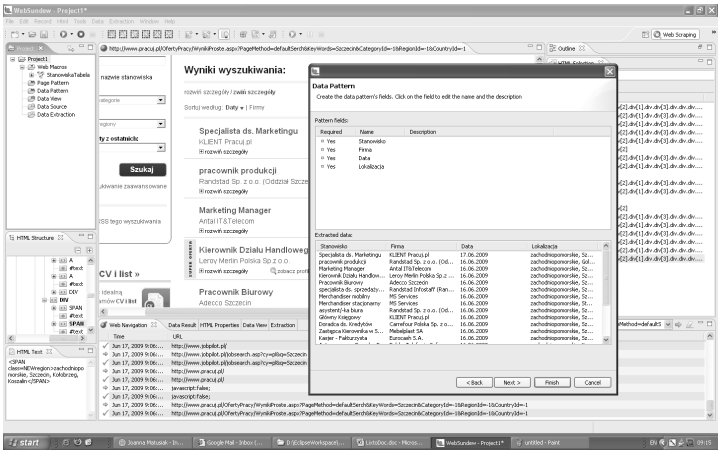


Rys. 6. Projektowanie wzorca danych – proces selekcji i identyfikacji danych
Źródło: opracowanie własne.



Rys. 7. Projektowanie wzorca danych – proces wyszukiwania wzorca

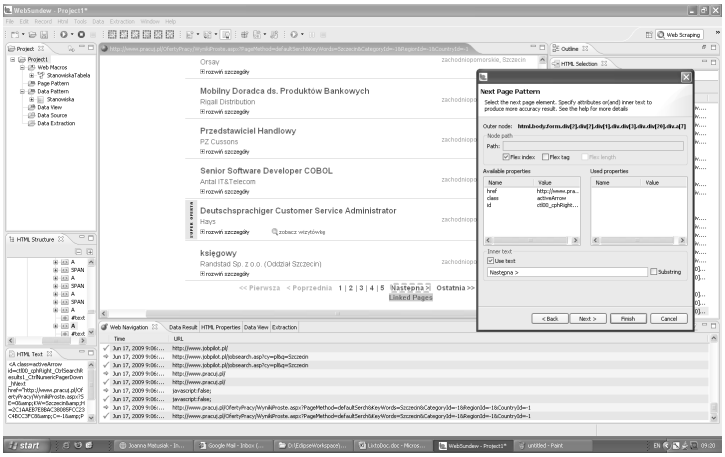
Źródło: opracowanie własne.



Rys. 8. Projektowanie wzorca podstron – proces weryfikacji wzorca na podstawie analizy otrzymanych danych z procesu ekstrakcji

Źródło: opracowanie własne.

6. Wzorzec kolejnej strony danych



Rys. 9. Projektowanie wzorca podstron – proces aktywacji wybranego typu wzorca i aktywacji „trybu pracy" nad stroną

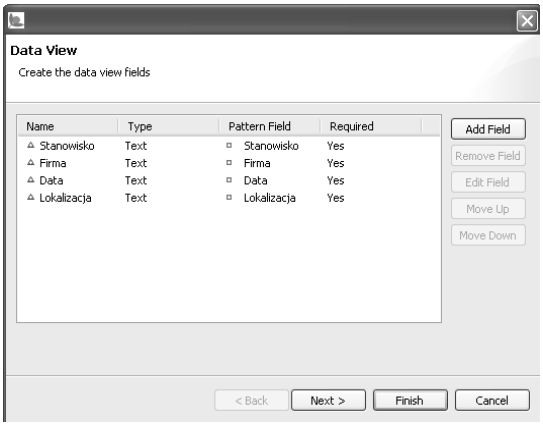
Źródło: opracowanie własne.

Rys. 10. Projektowanie wzorca kolejnych podstron – identyfikacja przejścia pomiędzy stronami

Źródło: opracowanie własne.

7. Widok danych

Widok danych odpowiada za skompletowanie pól danych w celu przygotowania ekstrakcji danych w kolejnym kroku. Każdemu polu danych (zmiennej identyfikacyjnej) musi zostać przypisane odpowiadające pole wzorca (rys. 11).

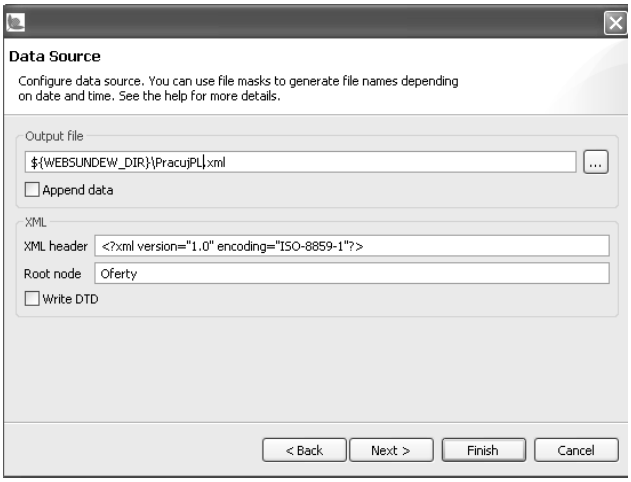


Rys. 11. Projektowanie widoku danych – pola danych i odpowiadające im pola wzorców
Źródło: opracowanie własne.

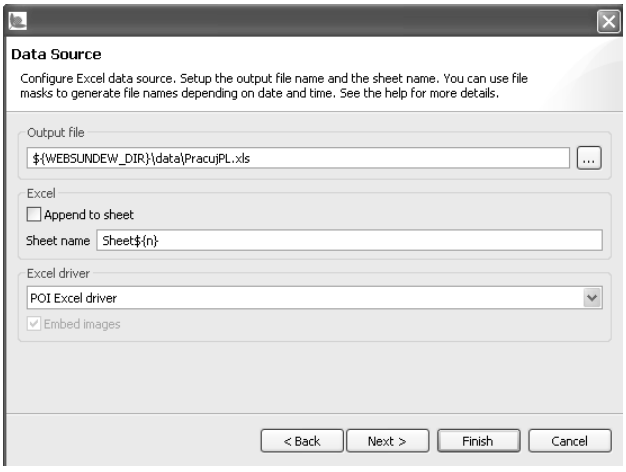
8. Konfiguracja źródeł wymiany danych

Aby zachować wynikowe dane procesu ekstrakcji, należy skonfigurować przynajmniej jedno źródło przechowywania danych. Narzędzie WebSundew oferuje następujące formaty plików do wyboru: XML, XLS, Text.

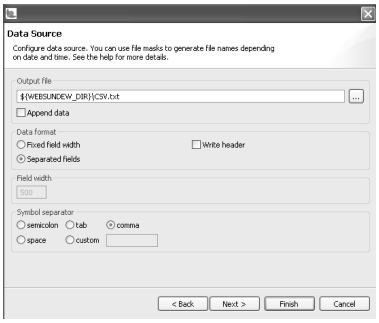
Na rysunku 12 przedstawiono konfigurację źródła przechowywania w pliku o formacie XML. Za pomocą tego kreatora użytkownik wskazuje nazwę macierzystego elementu XML (**Root Node**). W przypadku tekstowego przechowywania danych przedstawionego na rysunku 14, użytkownik może wybrać najodpowiedniejszą dla siebie formę zapisu, na przykład odseparowania danych (przy zaznaczonej opcji **Separated Fields**) – przez spację, średnik, przecinek itp.



Rys. 12. Konfiguracja źródła przechowywania danych – format XML
Źródło: opracowanie własne.



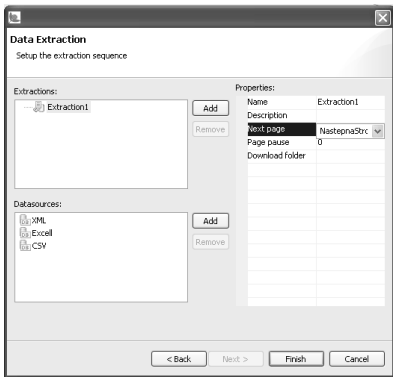
Rys. 13. Konfiguracja źródła przechowywania danych – format XLS
Źródło: opracowanie własne.



Rys. 14. Konfiguracja źródła przechowywania danych – format CSV
Źródło: opracowanie własne.

9. Przygotowanie ekstrakcji danych

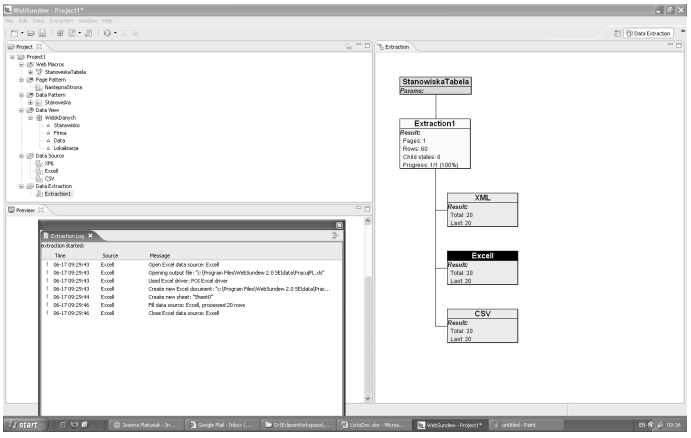
W celu przygotowania ekstrakcji danych należy wybrać utworzone źródło przechowywania danych, podać nazwę i opis ekstrakcji, a następnie nazwę utworzonego wzorca kolejnych podstroi (rys. 15).



Rys. 15. Projektowanie sekwencji ekstrakcji – wskazanie źródeł zapisu danych i wzorca kolejnych podstron
Źródło: opracowanie własne.

10. Uruchomienie ekstrakcji danych

Kolejnym krokiem po zdefiniowaniu ekstrakcji danych i źródła zapisywania jest uruchomienie ekstrakcji. W momencie uruchomienia użytkownik zostaje automatycznie przełączony do perspektywy **Data Extraction.** W trakcie wykonywania ekstrakcji dostępne są trzy widoki wyświetlające obecny status wykonywania i postępu, to jest: **Extraction**, **Extraction Log**, **Preview** (rys. 16).



Rys. 16. Rezultaty ekstrakcji danych
Źródło: opracowanie własne.

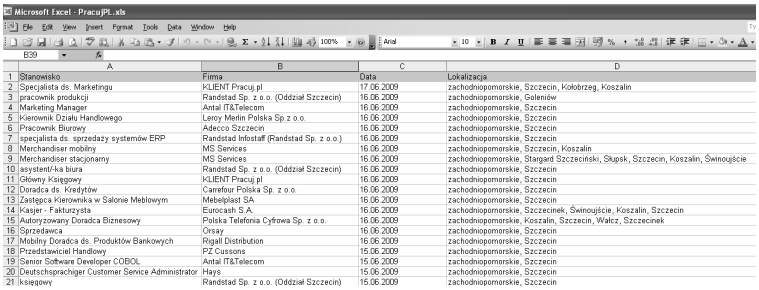
11. Uzyskane wyniki

W wyniku uruchomienia przygotowanej ekstrakcji danych, uzyskane dane w procesie ekstrakcji ofert pracy ogłoszonych na portalu **Pracuj.pl** dla kategorii wyszukiwania **„Szczecin"** zostały zapisane w trzech różnych formatach plików, dodanych uprzednio jako źródła przechowywania danych: XML, XLS i CSV. Zawartość tych plików przedstawiono na rysunkach 17 i 18.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<Oferty>
    <Pozycja>
        <Stanowisko>Doradca</Stanowisko>
        <Firma>LUKAS Bank SA</Firma>
        <Data>17.06.2009</Data>
        <Lokalizacja>zachodniopomorskie, Koszalin, Szczecin</Lokalizacja>
    </Pozycja>
    <Pozycja>
        <Stanowisko>Przedstawiciel Handlowy</Stanowisko>
        <Firma>KLIENT Pracuj.pl</Firma>
        <Data>17.06.2009</Data>
        <Lokalizacja>zachodniopomorskie, Szczecin, Kołobrzeg, Koszalin, Szczecinek, Drawsko Pomorskie, Czaplinek, Złocieniec, Sławno</Lokalizacja>
    </Pozycja>
    <Pozycja>
        <Stanowisko>Mobilny Doradca Bankowy</Stanowisko>
        <Firma>PBS Polska</Firma>
        <Data>17.06.2009</Data>
        <Lokalizacja>zachodniopomorskie, Szczecin</Lokalizacja>
    </Pozycja>
    <Pozycja>
        <Stanowisko>Specjalista ds. Marketingu</Stanowisko>
        <Firma>KLIENT Pracuj.pl</Firma>
        <Data>17.06.2009</Data>
        <Lokalizacja>zachodniopomorskie, Szczecin, Kołobrzeg, Koszalin</Lokalizacja>
    </Pozycja>
    <Pozycja>
        <Stanowisko>pracownik produkcji</Stanowisko>
        <Firma>Randstad Sp. z o.o. (Oddział Szczecin)</Firma>
        <Data>16.06.2009</Data>
        <Lokalizacja>zachodniopomorskie, Goleniów</Lokalizacja>
    </Pozycja>
    <Pozycja>
        <Stanowisko>Marketing Manager</Stanowisko>
        <Firma>Antal IT&Telecom</Firma>
        <Data>16.06.2009</Data>
        <Lokalizacja>zachodniopomorskie, Szczecin</Lokalizacja>
    </Pozycja>
    <Pozycja>
        <Stanowisko>Kierownik Działu Handlowego </Stanowisko>
        <Firma>Leroy Merlin Polska Sp.z o.o.</Firma>
```

Rys. 17. Wyniki ekstrakcji stanowisk z portalu pracy Pracuj.pl uzyskane dla kategorii przeszukiwania „Szczecin", zachowane w formacie pliku XLS

Źródło: opracowanie własne.



Microsoft Excel - PracujPL.xls

	A	B	C	D
1	Stanowisko	Firma	Data	Lokalizacja
2	Specjalista ds. Marketingu	KLIENT Pracuj.pl	17.06.2009	zachodniopomorskie, Szczecin, Kołobrzeg, Koszalin
3	pracownik produkcji	Randstad Sp. z o.o. (Oddział Szczecin)	16.06.2009	zachodniopomorskie, Goleniów
4	Marketing Manager	Antal IT&Telecom	16.06.2009	zachodniopomorskie, Szczecin
5	Kierownik Działu Handlowego	Leroy Merlin Polska Sp.z o.o.	16.06.2009	zachodniopomorskie, Szczecin
6	Pracownik Biurowy	Adecco Szczecin	16.06.2009	zachodniopomorskie, Szczecin
7	specjalista ds. sprzedaży systemów ERP	Randstad Infostaff (Randstad Sp. z o.o.)	16.06.2009	zachodniopomorskie, Szczecin
8	Merchandiser mobilny	MS Services	16.06.2009	zachodniopomorskie, Szczecin, Koszalin
9	Merchandiser stacjonarny	MS Services	16.06.2009	zachodniopomorskie, Stargard Szczeciński, Słupsk, Szczecin, Koszalin, Świnoujście
10	asystent/-ka biura	Randstad Sp. z o.o. (Oddział Szczecin)	16.06.2009	zachodniopomorskie, Szczecin
11	Główny Księgowy	KLIENT Pracuj.pl	16.06.2009	zachodniopomorskie, Szczecin
12	Doradca ds. Kredytów	Carrefour Polska Sp. z o.o.	16.06.2009	zachodniopomorskie, Szczecin
13	Zastępca Kierownika w Salonie Meblowym	Mebelplast SA	16.06.2009	zachodniopomorskie, Szczecin
14	Kasjer - Fakturzysta	Eurocash S.A.	16.06.2009	zachodniopomorskie, Szczecinek, Świnoujście, Koszalin, Szczecin
15	Autoryzowany Doradca Biznesowy	Polska Telefonia Cyfrowa Sp. z o.o.	16.06.2009	zachodniopomorskie, Koszalin, Szczecin, Wałcz, Szczecinek
16	Sprzedawca	Orsay	16.06.2009	zachodniopomorskie, Szczecin
17	Mobilny Doradca ds. Produktów Bankowych	Rigall Distribution	16.06.2009	zachodniopomorskie, Szczecin
18	Przedstawiciel Handlowy	PZ Cussons	15.06.2009	zachodniopomorskie, Szczecin
19	Senior Software Developer COBOL	Antal IT&Telecom	15.06.2009	zachodniopomorskie, Szczecin
20	Deutschsprachiger Customer Service Administrator	Hays	15.06.2009	zachodniopomorskie, Szczecin
21	księgowy	Randstad Sp. z o.o. (Oddział Szczecin)	15.06.2009	zachodniopomorskie, Szczecin

Rys. 18. Wyniki ekstrakcji stanowisk z portalu pracy Pracuj.pl uzyskane dla kategorii przeszukiwania „Szczecin", zachowane w formacie pliku XLS

Źródło: opracowanie własne.

Podsumowanie

W artykule przedstawiono możliwości praktycznego wykorzystania narzędzia do ekstrakcji zawartości stron internetowych w celu agregacji danych do dalszych analiz. Wskazano formaty wymiany danych: XML, XLS i CSV, dzięki którym pozyskane dane mogą stać się danymi wejściowymi złożonych systemów analitycznych. Dzięki wykorzystaniu technologii, takich jak eksploracja danych i ETL, systemy te mogą reorganizować dane, przeszukiwać je i w efekcie wyświetlać rezultaty analiz w formie raportów, wykresów i statystyk.

Literatura

http://web-harvest.sourceforge.net.
http://www.sundewsoft.com.
http://www.lixto.com.
http://www.newprosoft.com.
http://www.pracuj.pl.
http://www.eclipse.org.
http://www.eclipse.org/PDE.

EXTRACTION AND AGGREGATION OF THE JOB MARKET WEB SITES CONTENT

Summary

The article presents overview and practical exploration of the data extraction scraping tool for internet web sites content. As the exemplary analytical data source author has chosen job market portals offering the advertisements of new vacancies. Outcome results can be used in further detailed analysis as the input data of the complex analytical systems based on the data exploration, displaying search results according to the chosen criteria. Extraction data tool let the user store output results and exchange the data with other systems through XML, XSL and CSV files. Web scraping mechanism built into the tool offers graphical, action-based, user interactive processes. Data extraction is based on the web macro recordings as well as data and pages patterns generation.

Keywords: data extraction, data aggregation, job portals offers, job offers analysis

Translated by Joanna Matusiak