PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

**Mariusz Grabowski, Paweł Lula**
Uniwersytet Ekonomiczny w Krakowie

# EKSPLORACYJNA ANALIZA OFERT Z RYNKU NIERUCHOMOŚCI

**Streszczenie:** Zasadniczym celem pracy jest przedstawienie metody pozyskiwania informacji z nieustrukturyzowanych tekstowych ofert sprzedaży mieszkań. Zrealizowany proces badawczy składał się z następujących etapów: pozyskanie tekstów ofert z branżowego serwisu WWW, pozyskanie informacji za pomocą reguł zdefiniowanych w języku JAPE, przekształcenie pozyskanych zapisów do postaci tabelarycznej, realizacja obliczeń. Uzyskane w ten sposób informacje posłużyły do przeprowadzenia analizy krakowskiego rynku mieszkaniowego w 2009 r.

## 1. Wstęp

Istotność rynku nieruchomości jest trudna do przecenienia. Dla wielu jest miejscem dającym możliwość zaspokojenia jednaj z najważniejszych potrzeb – potrzeby posiadania domu czy mieszkania, przez innych rozpatrywany jest on jako miejsce lokowania kapitału, są również tacy, którzy traktują rynek nieruchomości jako barometr gospodarki. Niniejsza praca poświęcona jest badaniom jednego segmentu rynku nieruchomości, jakim jest rynek mieszkaniowy. Zakres terytorialny badań ograniczono do Krakowa. Natomiast jako źródło informacji przyjęto oferty sprzedaży mieszkań zamieszczone w jednym z branżowych portali internetowych.

## 2. Problematyka pozyskiwania informacji

Tematyka pozyskiwania informacji jest szeroko omawiana w literaturze. Bazując na pracy [Manning, Raghavan, Schütze 2009], można sformułować następującą definicję:

Termin „pozyskiwanie informacji" (*information retrieval*) odnosi się do metod i technik pozwalających na przeszukiwanie dużej kolekcji nieustrukturyzowanych zasobów informacyjnych w celu odnalezienia tych elementów, które spełniają przyjęte kryterium wyszukiwania.

Wskazane w powyższej definicji „nieustrukturyzowane zasoby informacyjne" odnoszą się zwykle do zasobów tekstowych w postaci plików tekstowych lub zawartości serwisów WWW.

Biorąc pod uwagę przeznaczenie pozyskanej informacji, można wskazać dwie typowe sytuacje:

a) wynik zrealizowanego procesu pozyskiwania informacji (czyli zbiór dokumentów zawierających informacje zgodne z wyspecyfikowanym kryterium) poddawany jest dalszemu przetworzeniu przez człowieka, który odpowiedzialny jest za właściwe zinterpretowanie i wykorzystanie znalezionych zapisów;

b) wynik zrealizowanego procesu pozyskiwania informacji poddawany jest przetworzeniu przez system komputerowy. Wydaje się, że w takim przypadku powyższa definicja nie obejmuje wszystkich koniecznych do zrealizowania elementów i wymaga przeformułowania.

Termin „pozyskiwanie informacji" (*information retrieval*) odnosi się do metod i technik pozwalających na:

- przeszukiwanie dużej kolekcji nieustrukturyzowanych zasobów informacyjnych,
- odnalezienie tych jej elementów, które spełniają przyjęte kryterium wyszukiwania,
- przekształcenie zapisów zawartych w zidentyfikowanych fragmentach do postaci ustrukturyzowanej, dogodnej do dalszego przetworzenia w sposób automatyczny.

## 3. Procedura badawcza zastosowana w badaniach krakowskiego rynku nieruchomości mieszkaniowych

Przeprowadzone badania dotyczyły krakowskiego rynku mieszkaniowego i przeprowadzono je w sierpniu i we wrześniu 2009 r. Jako materiał badawczy wykorzystano oferty sprzedaży mieszkań umieszczone w portalu oferty.net.


**Rys. 1.** Wyszukiwanie ofert w serwisie oferty.net

Źródło: opracowanie własne.

Zamieszczane w portalu oferty mają postać tekstową, częściowo ustrukturyzowaną. W każdym przypadku dwa pierwsze wiersze tekstu zawierają informację o lokalizacji mieszkania oraz jego cenie. Pozostałe fragmenty opisu nie mają już

jednolitego charakteru – w niektórych przypadkach umieszczony został fragment zawierający najważniejsze cechy mieszkania opisywane zgodnie ze schematem:

```
nazwa cechy: wartość cechy.
```

Następnie przedstawiany był tekst zawierający opis mieszkania.

Pierwszym etapem badań było utworzenie bazy ofert zawierającej opisy mieszkań w formacie tekstowym. W trakcie prac zastosowano następującą procedurę:

- Do pozyskania tekstów ofert zamieszczanych w serwisie oferty.net wykorzystano samodzielnie skonstruowanego pająka sieciowego przechodzącego po stronach serwisu i pobierającego strony HTML zawierające oferty sprzedaży mieszkań w Krakowie. Program ten zaimplementowany został w języku Java.
- Każda z pozyskanych stron zawierała elementy zbędne z punktu widzenia przeprowadzanej analizy (np. reklamy). Zachodziła więc potrzeba pozyskania wyłącznie istotnych fragmentów i ich przetworzenia z formatu HTML do postaci tekstowej. Ten etap analizy przeprowadzany był bezpośrednio po pobraniu strony. Przy jego implementacji wykorzystano parser: Jericho HTML (http://jericho.htmlparser.net/docs/index.html).
- Do ujednolicenia sposobu kodowania wykorzystano program Gżegżółka (http://www.gzegzolka.com/). Wszystkie dokumenty zostały przekonwertowane do formatu UTF-8.

Realizacja powyższej procedury pozwoliła na uzyskanie tekstowych wersji w postaci 10697 ofert. Tekst zawierający jedną z nich przedstawiony został poniżej:

*Kraków, Ruczaj-zaborze, Zalesie,*
*cena: 355 000 PLN (6920 PLN/m2),*
*Ulica: Zalesie,*
*Piętro: parter,*
*Liczba kondygnacji: 4,*
*Typ kuchni: do własnej aranżacji, jasna, oddzielna,*
*Hipoteczne; Czynsz: 250.00 zł; Budynek: blok, cegła, nowe budownictwo nowy; Standard mieszkania: do wprowadzenia; Dodatkowo: garderoba, nie ma piwnicy, balkon, drzwi antywłamaniowe, winda, teren ogrodzony, domofon; W pobliżu: sklepy, usługi, basen, fitness, kościół, przedszkole, szkoła, tereny rekreacyjne, Uniwersytet Jagielloński; Rozkład: do własnej aranżacji, ustawne, dwustronny, jasny, korzystny układ, pokoje nieprzechodnie; Ogrzewanie: centralne własne w budynku; Mieszkanie 2 pokoje nowe 51,3 $m^2$ , wykończone, Ruczaj ul. Zalesie od ulicy Zachodniej. Mieszkanie na parterze w czteropiętrowym bloku. Pokoje 14 $m^2$, 12 $m^2$, kuchnia 8 $m^2$, łazienka 6,5 $m^2$. W przedpokoju miejsce na garderobę, kuchnia w końcowej wersji z umeblowaniem. Mieszkanie ekonomiczne, własna kotłownia, baterie słoneczne na dachu – małe opłaty za ciepłą wodę.*

Drugi etap analizy miał na celu przekształcenie uzyskanych tekstów do postaci dogodnej do dalszego przetworzenia za pomocą metod statystycznych. Za-

stosowano podejście regułowe wymagające zdefiniowania wzorców określających sposób pozyskiwania informacji z tekstu i ich dalszego przetworzenia. Do zdefiniowania wzorców wykorzystano język JAPE pozwalający na definiowanie złożonych kryteriów wyszukiwania przy zastosowaniu wyrażeń regularnych [Thakker, Osman, Lakin 2009].

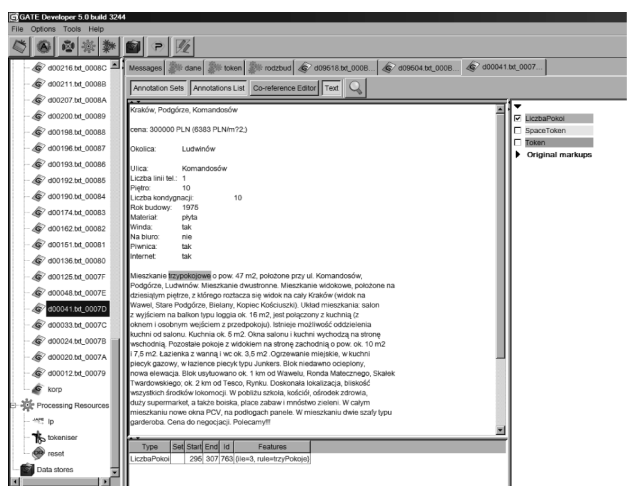
Ze względu na jednolity sposób podawania informacji o cenie mieszkania można w stosunkowo prosty i niezawodny sposób te informacje pozyskać z tekstu za pomocą reguły:

```
Phase: CenaMieszkania
Input: Token
Options: control = all
Rule: cenaMieszkania
(
   (
       {Token.string =~ "[Cc]ena"}
       {Token.kind == "punctuation"}
   )
   (
       {Token.kind == "number"}
   ):tempCalosc
   (
       {Token.kind == "word", Token.length ==3}
   ):tempJednostka
   (
       {Token.kind == "punctuation"}
   )
   (
       {Token.kind == "number"}
   ):tempCenam2
):cena
-->
:cena.Cena = {calosc = :tempCalosc.Token.string, cenam2 =
:tempCenam2.Token.string,jednostka = :tempJednostka.Token.string,
rule = cenaMieszkania}.
```

Wielość możliwości opisu pozostałych cech mieszkań znacznie utrudniała sformułowanie reguł pozwalających na pozyskanie innych, istotnych informacji o mieszkaniach. Tworzone w tym celu reguły były rozbudowane z powodu konieczności uwzględnienia wielu dopuszczalnych postaci opisów. Jako przykład może posłużyć fragment tekstu mający na celu identyfikację liczby pokoi występujących w mieszkaniu (jest to jedynie fragment mający na celu identyfikację zapisów mówiących o występowaniu dwóch pokoi):

```
Rule:dwaPokoje
 (
   (
       {Token.string =~ "[Dd]wupokojowe"}
   )
   |
   (
       {Token.string =~ "[Dd]w[au]"}
       {Token.string =~ "[Pp]oko[ij]"}
   )
   |
   (
       {Token.string =~ "2"}
       {Token.string =~ "pokoje"}
   )
):tempDwa
-->
:tempDwa.LiczbaPokoi = {ile = "2", rule = dwaPokoje}.
```

W trakcie badań wykorzystano wersję języka JAPE dostępną w systemie GATE (rys. 2).




**Rys. 2.** Realizacja zapytań w języku JAPE w systemie GATE

Źródło: opracowanie własne.

Realizacja poszczególnych reguł pozwoliła na przekształcenie informacji tekstowych do postaci tabelarycznej (rys. 3).

Wstępna analiza uzyskanych danych pozwoliła stwierdzić, że niektóre zapisy są zdublowane ze względu na zamieszczenie więcej niż jednej oferty dotyczącej tego samego mieszkania. Po usunięciu powielonych zapisów liczba ofert wykorzystanych w trakcie dalszej analizy wyniosła 9323.

Dane: dane po czyszczeniu.sta (8 zmn. * 9323 prz.)

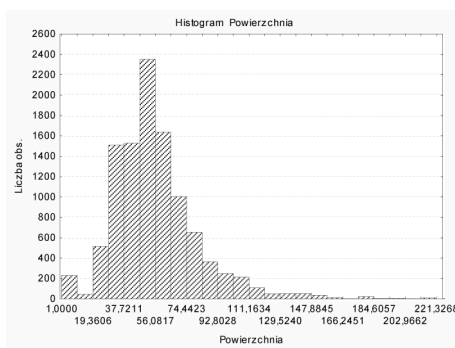
| | DANE | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 Plik | 2 Cena | 3 Cena-m2 | 4 Powierzchnia | 5 Kuchnia | 6 Liczba pokoi | 7 Piętro | 8 RodzajBud |
| 1 | d10697.tx | 10000000 | 36969 | 270,50 | oddzielna | 0 | 4 | kamienica |
| 2 | d10681.tx | 3000000 | 30000 | 100,00 | oddzielna | 0 | 1 | bd |
| 3 | d10693.tx | 6000000 | 27650 | 217,00 | oddzielna | 0 | 4 | bd |
| 4 | d10665.tx | 2500000 | 26882 | 93,00 | oddzielna | 2 | 2 | kamienica |
| 5 | d10691.tx | 5400000 | 25472 | 212,00 | polaczona | 0 | 2 | kamienica |
| 6 | d10636.tx | 2000000 | 24691 | 81,00 | oddzielna | 0 | 1 | bd |
| 7 | d10685.tx | 3327700 | 24334 | 136,75 | bd | 0 | 5 | bd |
| 8 | d10632.tx | 1950000 | 24122 | 80,84 | oddzielna | 2 | 1 | kamienica |
| 9 | d10673.tx | 2700000 | 23983 | 112,58 | aneks | 0 | 2 | kamienica |
| 10 | d10599.tx | 1700000 | 23866 | 71,23 | bd | 3 | 1 | apartamentowiec |
| 11 | d10669.tx | 2600000 | 23214 | 112,00 | aneks | 5 | 2 | kamienica |
| 12 | d10670.tx | 2600000 | 23214 | 112,00 | aneks | 0 | 2 | kamienica |
| 13 | d10398.tx | 1197700 | 23033 | 52,00 | aneks | 2 | 4 | apartamentowiec |
| 14 | d10503.tx | 1320000 | 22759 | 58,00 | oddzielna | 2 | 1 | kamienica |
| 15 | d10667.tx | 2523198 | 22713 | 111,09 | bd | 0 | 4 | kamienica |
| 16 | d10672.tx | 2640000 | 22000 | 120,00 | oddzielna | 3 | 2 | kamienica |
| 17 | d10686.tx | 3600000 | 21752 | 165,50 | oddzielna | 3 | 2 | kamienica |
| 18 | d10687.tx | 3600000 | 21687 | 166,00 | oddzielna | 3 | 2 | kamienica |
| 19 | d06816.tx | 420648 | 21451 | 19,61 | bd | 0 | 0 | kamienica |
| 20 | d10095.tx | 900000 | 21429 | 42,00 | oddzielna | 0 | 2 | kamienica |
| 21 | d10097.tx | 900000 | 21429 | 42,00 | aneks | 1 | 2 | kamienica |
| 22 | d10680.tx | 2889000 | 21400 | 135,00 | oddzielna | 0 | 3 | bd |
| 23 | d10054.tx | 890000 | 21190 | 42,00 | aneks | 0 | 2 | kamienica |
| 24 | d10677.tx | 2772000 | 21000 | 132,00 | aneks | 4 | 2 | kamienica |
| 25 | d07979.tx | 499000 | 20940 | 23,83 | aneks | 0 | 2 | kamienica |
| 26 | d07980.tx | 499000 | 20792 | 24,00 | aneks | 0 | 2 | kamienica |
| 27 | d10676.tx | 2737198 | 20683 | 132,34 | bd | 0 | 3 | kamienica |
| 28 | d10678.tx | 2778472 | 20642 | 134,60 | oddzielna | 0 | 5 | bd |

**Rys. 3.** Wyniki zapytań po przekształceniu do postaci tabelarycznej

Źródło: opracowanie własne.

## 4. Krakowski rynek nieruchomości mieszkaniowych w świetle przeprowadzonych badań

Rozkład zmiennej reprezentującej powierzchnię sprzedawanych mieszkań przedstawia wykres zamieszczony na rys. 4.

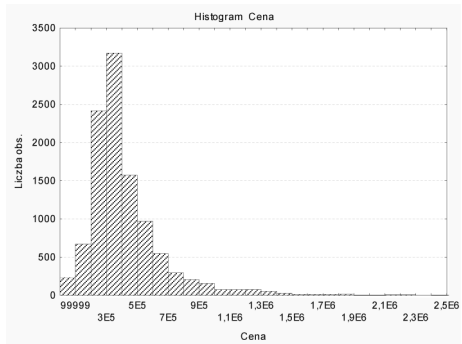



**Rys. 4.** Rozkład wartości reprezentujących powierzchnię mieszkań

Źródło: opracowanie własne.

Wartość średnia zmiennej *Powierzchnia* wyniosła 58,3 $m^2$, mediana zaś 53 $m^2$, co potwierdza tezę o szczególnie dużym udziale mieszkań średnich w obrocie.

Rozkład zmiennej *Cena mieszkania* przedstawiony został na rys. 5.

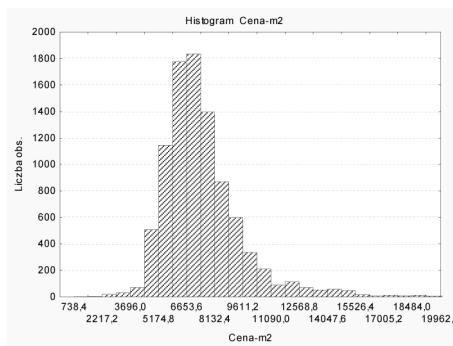



**Rys. 5.** Rozkład wartości reprezentujących cenę mieszkań

Źródło: opracowanie własne.

W przypadku zmiennej *Cena mieszkania* wartość średnia wyniosła 444,8 tys. zł, mediana zaś 365 tys. zł.

Analizowana była również zmienna *Cena jednego metra kwadratowego mieszkania*. Rozkład tej zmiennej przedstawia wykres zamieszczony na rys. 6.



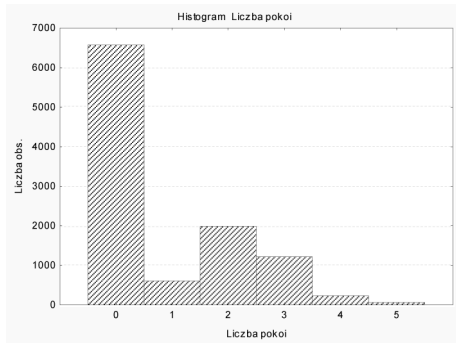

**Rys. 6.** Rozkład wartości reprezentujących cenę za jeden metr kw. mieszkania

Źródło: opracowanie własne.

Wartość średnia to 7515 zł, mediana zaś wyniosła 7061 zł.

Istotną cechą mieszkania jest liczba pokoi. Kształtowanie się tej wartości przedstawiono na rys. 7.

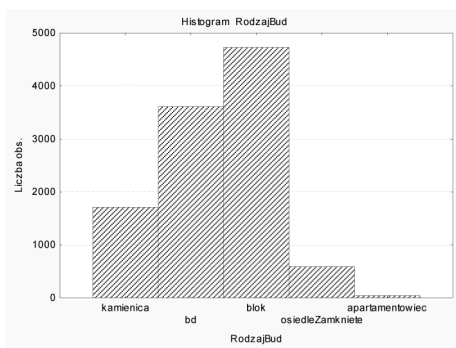
Zerowa liczba pokoi reprezentuje wszystkie te przypadki, w których zastosowane reguły nie zidentyfikowały odpowiedniej informacji w tekście (mogło to wynikać z niedoskonałości reguł lub z braku odpowiedniego zapisu w ofercie sprzedaży). Wśród przypadków, w których udało się pozyskać informację o liczbie pokoi, występuje wyraźna przewaga mieszkań dwupokojowych.




**Rys. 7.** Rozkład wartości reprezentujących liczbę pokoi w oferowanych mieszkaniach

Źródło: opracowanie własne.

Ciekawe wyniki daje również analiza informacji dotyczących *rodzaju budynku*, w którym zlokalizowane jest mieszkanie (rys. 8).




**Rys. 8.** Rozkład zmiennej *Rodzaj budynku*

Źródło: opracowanie własne.

Największy udział w obrocie mają mieszkania zlokalizowane w blokach mieszkalnych. Wyraźnie uwidacznia się kategoria „osiedla zamknięte". Jednocześnie należy zauważyć, że w stosunkowo dużej liczbie ofert nie udało się określić typu budynku.

## 5. Podsumowanie

Przeprowadzone badania pozwalają na sformułowanie kilku wniosków:

- ze względu na wysoki stopień upowszechnienia informacji tekstowej (zwłaszcza w Internecie) za szczególnie uzasadnione należy uznać prace zmierzające

do opracowywania i doskonalenia metod pozwalających na automatyzację procesu ich przetwarzania,

- zastosowana regułowa metoda pozyskiwania informacji z dokumentów tekstowych pozwoliła na uzyskanie stosunkowo dobrych rezultatów; próbując uogólnić te wyniki, należy podkreślić, że analizowany zbiór tekstów miał wyraźnie monotematyczny charakter; w przypadku analizy tekstów o większym zróżnicowaniu tematycznym zaproponowanie reguł byłoby bardzo utrudnione,
- przeprowadzając ocenę zastosowanej metody badawczej, należy podkreślić konieczność skorzystania z wiedzy eksperckiej na etapie konstruowania reguł oraz czasochłonność procesu analizy,
- wykonane analizy potwierdzają przydatność pakietu GATE. To udostępniane bezpłatnie narzędzie jest zaawansowane, a jednocześnie stabilne, wspomagając eksplorację zasobów tekstowych,
- zastosowane w pracy metody pozyskiwania informacji z dokumentów tekstowych pozwoliły na przeprowadzenie analizy ofert dotyczących krakowskiego rynku mieszkaniowego.

## Literatura

Lula P., *Text Mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych*, Statsoft, 2005, http://www.statsoft.pl/czytelnia/8_2007/Lula05.pdf.

Manning C.D., Raghavan P., Schütze H., *An Introduction to Information Retrieval*, Cambridge University Press, Cambridge, England 2009.

Thakker D., Osman T., Lakin P., *GATE JAPE Grammar Tutorial*, http://www.gate.ac.uk/sale/thakker-jape-tutorial/GATE%20JAPE%20manual.pdf, 2009.

## EXPLORATORY ANALYSIS OF SELL OFFERS IN REAL ESTATE MARKET

**Summary:** The paper discusses the problem of information retrieval from text documents containing sell offers concerning apartments in Cracow. The research process consists of several stages: offers' set preparing (corpus preparing), information retrieval from corpus by JAPE rules, conversion to table representation, data analysis. A short description on apartment market in Cracow was presented in the last part of the paper.