

**PRACE NAUKOWE**

Uniwersytetu Ekonomicznego we Wrocławiu

**RESEARCH PAPERS**

of Wrocław University of Economics

**242**

# **Taksonomia 19.**

## **Klasyfikacja i analiza danych – teoria i zastosowania**



Redaktorzy naukowi  
**Krzysztof Jajuga**  
**Marek Walesiak**



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu  
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pocięcha,  
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS  
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie [www.ibuk.pl](http://www.ibuk.pl)

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych  
The Central European Journal of Social Sciences and Humanities <http://cejsh.icm.edu.pl>  
oraz w The Central and Eastern European Online Library [www.ceeol.com](http://www.ceeol.com),  
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon [http://kangur.uek.krakow.pl/  
bazy\\_ae/bazekon/nowy/index.php](http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php)

Informacje o naborze artykułów i zasadach recenzowania znajdują się  
na stronie internetowej Wydawnictwa  
[www.wydawnictwo.ue.wroc.pl](http://www.wydawnictwo.ue.wroc.pl)

Kopiowanie i powielanie w jakiegokolwiek formie  
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu  
Wrocław 2012

**ISSN 1899-3192** (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)  
**ISSN 1505-9332** (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM  
Nakład: 320 egz.

## Spis treści

<b>Wstęp</b> .....	13
<b>Stanisława Bartosiewicz</b> , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej .....	17
<b>Andrzej Sokolowski</b> , Q uniwersalna miara odległości .....	22
<b>Eugeniusz Gatnar</b> , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP) .....	31
<b>Marek Walesiak</b> , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
<b>Krzysztof Jajuga, Marek Walesiak</b> , XXV lat konferencji taksonomicznych – fakty i refleksje .....	47
<b>Józef Pocięcha, Barbara Pawelek</b> , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne .....	50
<b>Paweł Lula</b> , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych .....	58
<b>Ewa Roszkowska</b> , Zastosowanie metody TOPSIS do wspomaganie procesu negocjacji.....	68
<b>Andrzej Młodak</b> , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne .....	76
<b>Andrzej Bąk</b> , Modele kategorii nieuporządkowanych w badaniach preferencji .....	86
<b>Jacek Kowalewski</b> , Zintegrowany model optymalizacji badań statystycznych.....	96
<b>Jan Paradysz, Karolina Paradysz</b> , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
<b>Tomasz Szubert</b> , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
<b>Izabela Szamrej-Baran</b> , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne .....	126
<b>Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski</b> , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
<b>Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz</b> , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych .....	144
<b>Hanna Dudek</b> , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów .....	153

<b>Joanicjusz Nazarko, Ewa Chodakowska, Marta Jaročka,</b> Segmentacja szkół wyższych metodą analizy skupień <i>versus</i> konkurencja technologiczna ustalona metodą DEA – studium komparatywne.....	163
<b>Ewa Chodakowska,</b> Wybrane metody klasyfikacji w konstrukcji ratingu szkół.....	173
<b>Bartosz Soliński,</b> Sektor energetyki odnawialnej w krajach Unii Europejskiej – klasyfikacja w świetle strategii zarządzania zmianą.....	182
<b>Krzysztof Szwarz,</b> Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną.....	192
<b>Elżbieta Gołata, Grażyna Dehnel,</b> Rejestry administracyjne w analizie przedsiębiorczości.....	202
<b>Katarzyna Chudy, Marek Sobolewski, Kinga Stępień,</b> Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce.....	212
<b>Katarzyna Dębowska,</b> Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod dyskryminacji i regresji.....	222
<b>Alina Bojan,</b> Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji.....	231
<b>Justyna Brzezińska,</b> Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE.....	240
<b>Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka,</b> Analiza klas ukrytych w badaniach satysfakcji studentów.....	247
<b>Bartłomiej Jefmański,</b> Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska R.....	256
<b>Julita Stańczuk,</b> Porównanie rezultatów wielostanowej klasyfikacji obiektów ekonomicznych z wykorzystaniem analizy dyskryminacyjnej oraz sieci neuronowych.....	265
<b>Jerzy Krawczuk,</b> Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500.....	275
<b>Anna Czapkiewicz, Beata Basiura,</b> Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych na podstawie modelu Copula-GARCH.....	283
<b>Radosław Pietrzyk,</b> Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych.....	291
<b>Aleksandra Witkowska, Marek Witkowski,</b> Zastosowanie metody Panzara-Rosse’a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych.....	306
<b>Marcin Pelka,</b> Podejście wielomodelowe z wykorzystaniem metody <i>boosting</i> w analizie danych symbolicznych.....	315
<b>Justyna Wilk,</b> Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych.....	323

<b>Tomasz Bartłomowicz, Justyna Wilk</b> , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
<b>Kamila Migdał-Najman</b> , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących .....	342
<b>Dorota Rozmus</b> , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i> .....	352
<b>Krzysztof Najman</b> , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG .....	361
<b>Małgorzata Misztal</b> , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna .....	370
<b>Mariusz Kubus</b> , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
<b>Barbara Batóg, Jacek Batóg</b> , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym .....	387
<b>Katarzyna Wójcik, Janusz Tuchowski</b> , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej .....	396
<b>Iwona Staniec</b> , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach .....	406
<b>Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawelczyk, Jerzy Kołodziej, Jerzy Błaszczyk</b> , Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami .....	416
<b>Iwona Foryś</b> , Wykorzystanie analizy log-liniowej do wyboru czynników determinujących atrakcyjność cenową mieszkań w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego.....	426
<b>Ewa Genge</b> , Analiza skupień oparta na mieszankach uciętych rozkładów normalnych.....	436
<b>Jerzy Korzeniewski</b> , Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień .....	444
<b>Andrzej Dudek</b> , SMS – propozycja nowego algorytmu analizy skupień .....	451
<b>Artur Mikulec</b> , Metody oceny wyniku grupowania w analizie skupień.....	460
<b>Małgorzata Machowska-Szewczyk</b> , Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych .....	469
<b>Artur Zaborski</b> , Analiza PROFIT i jej wykorzystanie w badaniu preferencji .....	479
<b>Karolina Bartos</b> , Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena .....	488

<b>Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak</b> , Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych .	496
<b>Izabela Kurzawa</b> , Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto–wieś .....	505
<b>Aleksandra Łuczak, Feliks Wysocki</b> , Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych .....	513
<b>Agnieszka Sompolska-Rzechuła</b> , Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim .....	523
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk</b> , Ocena intensywności wykorzystania skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego .....	532
<b>Iwona Bąk</b> , Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę .....	541
<b>Aneta Becker</b> , Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010	552
<b>Katarzyna Dębowska</b> , Klasyfikacja sektorów ze względu na ich kondycję finansową przy użyciu metod wielowymiarowej analizy statystycznej .....	562
<b>Anna Domagała</b> , Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru w przód).....	571
<b>Henryk Gierszal, Karina Pawlina, Maria Urbańska</b> , Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej .....	580
<b>Hanna Gruchociak</b> , Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej.....	590
<b>Tomasz Klimanek, Marcin Szymkowiak</b> , Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy .....	601
<b>Jarosław Lira</b> , Prognozowanie opłacalności produkcji żywca wieprzowego w Polsce .....	610
<b>Christian Lis</b> , Wykorzystanie metody klasyfikacji w ocenie konkurencyjności portów południowego Bałtyku .....	619
<b>Beata Bieszk-Stolorz, Iwona Markowicz</b> , Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych .	628
<b>Lucyna Przezbórska-Skobiej, Jarosław Lira</b> , Przestrzeń agroturystyczna Polski i ocena jej atrakcyjności.....	637
<b>Paweł Ulman</b> , Model rozkładu wydatków a funkcje popytu.....	646
<b>Maria Urbańska, Tadeusz Mizera, Henryk Gierszal</b> , Zastosowanie metod analizy statystycznej w badaniach mięczaków .....	655

## Summaries

<b>Stanisława Bartosiewicz</b> , The effects of subjectivism in multivariate analysis revisited.....	21
<b>Andrzej Sokółowski</b> , Q universal distance measure .....	30
<b>Eugeniusz Gatnar</b> , Data quality in central banks' statistical systems (NBP example) .....	38
<b>Marek Walesiak</b> , Distance measures for ordinal data – strategies of proceedings.....	46
<b>Krzysztof Jajuga, Marek Walesiak</b> , XXV years of taxonomic conferences – some facts and remarks.....	49
<b>Józef Pocięcha, Barbara Pawelek</b> , General SEM model in researching corporate bankruptcy and business cycles – theoretical and practical problems.....	57
<b>Paweł Lula</b> , Learning-based systems of information extraction from textual resources .....	67
<b>Ewa Roszkowska</b> , The application of the TOPSIS method to support the negotiation process .....	75
<b>Andrzej Młodak</b> , Neighborhood of spatial areas in the physical and socio-economic context – a taxonomic approach.....	85
<b>Andrzej Bąk</b> , Models for unordered categories in preference analysis.....	95
<b>Kowalewski Jacek</b> , An integrated model of optimizing statistical surveys ....	105
<b>Jan Paradysz, Karolina Paradysz</b> , Areas of unemployment in Poland – benchmark problem .....	115
<b>Tomasz Szubert</b> , How to play to lose the least? Classification of systems in sports bets .....	125
<b>Izabela Szamrej-Baran</b> , Classification of EU member states in view of fuel poverty .....	134
<b>Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski</b> , An attempt to use the gravity model in the analysis of commuters.....	143
<b>Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz</b> , Subsistence minimum versus factors influencing tendency to benefit from social care. Classification of households .....	152
<b>Hanna Dudek</b> , Subjective equivalence scales – analysis based on data about satisfaction with incomes.....	162
<b>Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka</b> , Segmentation of universities using cluster analysis versus technological competitors determined by the DEA method – a comparative study .....	172
<b>Ewa Chodakowska</b> , Selected methods of classification in schools' rating.....	181
<b>Bartosz Soliński</b> , Renewable energy sector in the European Union – classification in the light of change management strategy .....	191
<b>Krzysztof Szwarc</b> , Classification of Wielkopolska voivodeship due to the demographic situation .....	201

<b>Elżbieta Gołata, Grażyna Dehnel</b> , Administrative registers in business analysis.....	211
<b>Katarzyna Chudy, Marek Sobolewski, Kinga Stępień</b> , Application of taxonomic methods in forecasting the profitability ratios of listed banks in Poland.....	221
<b>Katarzyna Dębowska</b> , Modeling bankruptcy of firms by using discrimination and regression methods.....	230
<b>Alina Bojan</b> , Identification of variables which influence attractiveness of given investments with the usage of multivariate analysis.....	239
<b>Justyna Brzezińska</b> , Log-linear analysis in the study of mortality in EU.....	246
<b>Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka</b> , Latent class analysis in student satisfaction surveys.....	254
<b>Bartłomiej Jefmański</b> , The respondent's opinions measurement in the R program with an application of fuzzy sets theory.....	264
<b>Julita Stańczuk</b> , A comparison of the results of multistate classification of economic objects using discriminant analysis and artificial neural networks.....	274
<b>Jerzy Krawczuk</b> , Effectiveness of classification methods in S&P500 stock index direction changes forecasting.....	282
<b>Anna Czapkiewicz, Beata Basiura</b> , The simulation study of the utility of the Copula-GARCH models for clustering financial time series.....	290
<b>Radosław Pietrzyk</b> , Timing and selectivity in mutual funds performance measurement.....	305
<b>Aleksandra Witkowska, Marek Witkowski</b> , Use of the Panzar-Rosse method to assess of the competition level in the cooperative banks sector.....	314
<b>Marcin Pelka</b> , Ensemble learning with the application of <i>boosting</i> in symbolic data analysis.....	322
<b>Justyna Wilk</b> , Comparative study of symbolic data classification software.....	332
<b>Tomasz Bartłomowicz, Justyna Wilk</b> , Application of symbolic data analysis methods for domain database searching.....	341
<b>Kamila Migdał-Najman</b> , A proposal of hybrid clustering method based on self-learning networks.....	351
<b>Dorota Rozmus</b> , Comparison of accuracy of spectral clustering and cluster ensembles stability based on bagging idea.....	360
<b>Krzysztof Najman</b> , A dynamic grouping based on self-learning GNG networks.....	369
<b>Małgorzata Misztal</b> , Influence of data imputation methods on the results of object classification using classification trees in the case of small data sets – simulation assessment.....	379
<b>Mariusz Kubus</b> , The application of pre-conditioning of explanatory variable for feature selection.....	386
<b>Barbara Batóg, Jacek Batóg</b> , Application of discriminant analysis to the identification of factors determining the rate of return on the capital market.....	395



<b>Katarzyna Wójcik, Janusz Tuchowski</b> , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
<b>Iwona Staniec</b> , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
<b>Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk</b> , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
<b>Iwona Foryś</b> , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
<b>Ewa Genge</b> , Trimming approach to the mixtures of normal distributions.....	443
<b>Jerzy Korzeniewski</b> , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
<b>Andrzej Dudek</b> , SMS – proposal of new clustering algorithm.....	459
<b>Artur Mikulec</b> , Evaluation methods for the grouping result in cluster analysis.....	468
<b>Małgorzata Machowska-Szewczyk</b> , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
<b>Artur Zaborski</b> , PROFIT analysis and its using in the research of preferences.....	487
<b>Karolina Bartos</b> , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
<b>Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak</b> , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
<b>Izabela Kurzawa</b> , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
<b>Aleksandra Luczak, Feliks Wysocki</b> , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
<b>Agnieszka Sompolska-Rzechuła</b> , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk</b> , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
<b>Iwona Bąk</b> , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
<b>Aneta Becker</b> , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
<b>Katarzyna Dębowska</b> , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

---

<b>Anna Domagała</b> , Proposal of a new method for variable selection in DEA models (combined forward stepwise selection method).....	579
<b>Henryk Gierszal, Karina Pawlina, Maria Urbańska</b> , Statistical analysis in demand research of ICT services in mobile networks.....	589
<b>Hanna Gruchociak</b> , Construction of regression estimator for two-level data	600
<b>Tomasz Klimanek, Marcin Szymkowiak</b> , Application of spatial models in indirect estimation of some labor market characteristics .....	609
<b>Jarosław Lira</b> , Forecasting of hog livestock production profitability in Poland .....	618
<b>Christian Lis</b> , The utilization of taxonomic methods in the appraisal of competitiveness of south Baltic ports .....	627
<b>Beata Bieszk-Stolorz, Iwona Markowicz</b> , The application of the multinomial logit model in evaluating employment odds for the unemployed job seekers .....	636
<b>Lucyna Przezbórska-Skobiej, Jarosław Lira</b> , Agritourism space of Poland and its valuation.....	645
<b>Paweł Ulman</b> , Model of expenses distribution and demand functions.....	654
<b>Maria Urbańska, Tadeusz Mizera, Henryk Gierszal</b> , Methods of statistical analysis in research of molluscs .....	663

**Paweł Lula**

Uniwersytet Ekonomiczny w Krakowie

---

## UCZĄCE SIĘ SYSTEMY POZYSKIWANIA INFORMACJI Z DOKUMENTÓW TEKSTOWYCH

---

**Streszczenie:** Zasadniczym celem pracy jest prezentacja, klasyfikacja i ocena systemów pozyskiwania informacji z dokumentów tekstowych konstruowanych przy udziale mechanizmów uczących. W początkowej części pracy zdefiniowano pojęcie systemu uczącego się oraz zagadnienie pozyskiwania informacji. Następnie zaprezentowano strukturę oraz sposób funkcjonowania uczącego się systemu pozyskiwania informacji. Kluczowym elementem tego typu rozwiązań jest model zawartości informacyjnej. Jego charakterystyka i rodzaje są zasadniczym tematem kolejnego punktu pracy. Następnie przedstawiono rolę wiedzy zewnętrznej i metod uczenia maszynowego w poszczególnych rozwiązaniach. W kolejnej części artykułu zaprezentowano rozważania dotyczące lokalnego lub globalnego charakteru poszczególnych rozwiązań.

**Słowa kluczowe:** pozyskiwanie informacji z dokumentów tekstowych, analiza tekstu, text mining.

### 1. Wstęp

Występujący praktycznie w każdej dziedzinie dynamiczny wzrost liczby dokumentów sprawia, że koniecznością staje się automatyzowanie procesów pozyskiwania informacji z dokumentów tekstowych. W wielu przypadkach konstrukcja tego typu rozwiązań jest zadaniem praco- i czasochłonnym. Dlatego też pozytywnie należy ocenić podejmowane próby zastąpienia zaangażowania człowieka w budowę systemu przez mechanizmy pozwalające na doskonalenie systemu dzięki zastosowaniu algorytmów uczenia maszynowego.

Zasadniczym celem pracy jest prezentacja, klasyfikacja i ocena systemów pozyskiwania informacji z dokumentów tekstowych konstruowanych przy udziale mechanizmów uczących.

### 2. System uczący się

System uczący się charakteryzuje się zdolnością do poprawy sposobu swojego funkcjonowania na podstawie wniosków płynących z przeprowadzonej przez niego ana-

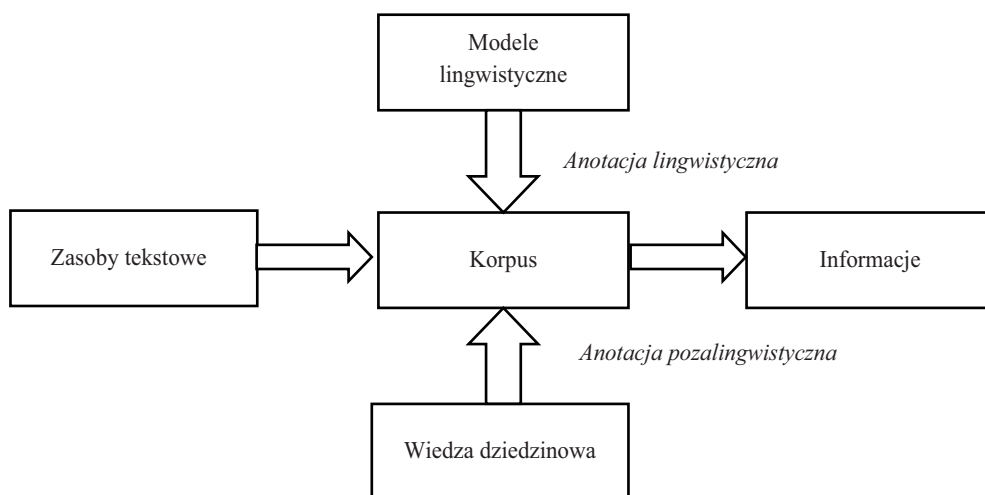
lizey dostarczonych danych. W przeważającej liczbie przypadków systemy uczące się przyjmują postać algorytmów komputerowych służących do rozwiązywania ściśle określonej klasy zadań. Niech  $P$  będzie przyjętą miarą jakości sposobu realizacji rozpatrywanego typu zadania, zaś  $D$  – zbiorem danych opisującym rzeczywisty przebieg procesów zmierzających do ich rozwiązania. Algorytmowi  $A$  można przypisać zdolność do *uczenia się*, jeśli został on przystosowany do polepszania swojego działania w sensie przyjętej miary  $P$  poprzez analizę danych zawartych w zbiorze  $D$ .

### 3. Pozyskiwanie informacji

Pozyskiwanie informacji z zasobów tekstowych będziemy rozumieć jako obszar badań naukowych oraz zastosowań wypracowanych metod i narzędzi mających na celu wydobycie istotnych informacji z zasobów przyjmujących postać nieustrukturyzowanego tekstu (np. dokumentów w postaci elektronicznej, zawartości stron WWW, blogów, poczty elektronicznej). System pozyskiwania informacji wykonuje swoje zadanie jako odpowiedź na zapytanie sformułowane przez użytkownika.

### 4. Struktura i etapy budowy uczącego się systemu pozyskiwania informacji z zasobów tekstowych

Struktura systemu pozyskiwania informacji z zasobów tekstowych przedstawiona została na rys. 1.



**Rys. 1.** Struktura systemu pozyskiwania informacji

Źródło: opracowanie własne.

Punktem wyjścia w procesie budowy uczącego się systemu pozyskiwania informacji z dokumentów tekstowych jest przekształcenie zestawu dokumentów źródłowych w jednolity korpus. Operacja ta obejmuje przekształcenie dokumentów do postaci plików tekstowych, usunięcie wszelkich informacji formatujących oraz ujednolicenie sposobu kodowania.

Kolejnym etapem jest anotacja (*annotation*) korpusu. Pojęcie anotacji określane jest również mianem znakowania. Polega ona na dodaniu do tekstu źródłowego dodatkowych informacji w postaci znaczników obejmujących swoim zasięgiem fragmenty tekstu. Znaczniki odgrywają zwykle podwójną rolę: określają charakter fragmentu tekstu przez zidentyfikowanie jego charakteru i przypisanie mu właściwej etykiety oraz pozwalają na przypisanie pewnych dodatkowych informacji do fragmentów tekstu źródłowego. Jako przykład ilustrujący wspomniane dwie funkcje posłużyć może zapis:

**Jacek<CZASOWNIK bezokolicznik="kolekcjonować">kolekcjonuje</CZASOWNIK>znaczki pocztowe.**

Dodane znaczniki wskazują, że słowo „kolekcjonuje” jest czasownikiem i jednocześnie informują o postaci bezokolicznika.

Anotacja może dotyczyć informacji lingwistycznej lub pozalingwistycznej. Znaczniki pozalingwistyczne mają zwykle na celu wzbogacenie warstwy znaczeniowej tekstu, dlatego też określane są mianem anotacji semantycznej.

Najbardziej upowszechnioną formą anotacji lingwistycznej jest identyfikacja części mowy (*POS tagging* lub *POST – part-of-speech tagging*). W trakcie realizacji tego zadania poszczególnym wyrazom przypisywane są informacje określające m.in. właściwą dla niego część mowy, postać podstawową, formę występującą w dokumencie. Identyfikacja części mowy może być realizowana za pomocą słowników lub odpowiednich algorytmów.

Natomiast powszechnie stosowanym elementem anotacji pozalingwistycznej jest identyfikacja znaczeniowa elementów tekstu (*NER – Named Entity Recognition*). Zadanie polega na wyodrębnieniu takich elementów, jak: imiona i nazwiska osób, nazwy geograficzne, nazwy organizacji i instytucji, określenia czasu, jednostki pieniężne, adresy poczty elektronicznej, numery telefonów i wiele innych. Elementy te są następnie znakowane.

Korpus dokumentów wzbogacony o dodatkowe informacje dodane w postaci anotacji określane jest mianem *korpusu anotowanego*. Jest on punktem wyjścia do budowy *modelu zawartości informacyjnej* korpusu. Podstawowym zadaniem modelu jest reprezentowanie informacji występujących w tekście, zapewnienie do nich dostępu oraz dostarczenie narzędzi pozwalających na ich przetwarzanie w sposób pożądaný przez użytkownika systemu.

## 5. Modele zawartości informacyjnej dokumentów

Prezentowane w niniejszej pracy podejście zakłada, że system pozyskiwania informacji nie operuje bezpośrednio na dokumentach źródłowych, lecz na modelu reprezentującym informacje pochodzące z dokumentów. Przyjęty sposób modelowego ujęcia informacji ma istotny wpływ na możliwości i efektywność systemu. W sposób znaczący determinuje również koszty związane z budową systemu. W niniejszej pracy wyróżnione zostały cztery typy modeli opisujących zawartość dokumentów. Każdy z nich może zostać wykorzystany w systemach pozyskiwania informacji. Jednakże funkcje, jakie mogą spełniać poszczególne rozwiązania, są bardzo zróżnicowane.

## 6. Modele oparte na koncepcji przestrzeni wektorowej

W przedstawianym podejściu konstruowana jest przestrzeń, której poszczególne wymiary odpowiadają terminom indeksującym przyjętym dla danego zestawu dokumentów. W najprostszym przypadku funkcję terminów indeksujących spełniać mogą wyrazy (ewentualnie po sprowadzeniu do formy podstawowej). W bardziej zaawansowanych podejściach rolę terminów indeksujących mogą odgrywać frazy lub identyfikatory przypisane do zidentyfikowanych w tekście faktów. Dokument jest więc kombinacją liniową wektorów odpowiadających występującym w nim terminów indeksujących. Bardzo ważnym problemem jest sposób ustalania wag odpowiadających poszczególnym wymiarom. Najczęściej spotykaną metodą jest technika *tf-idf* zakładająca, że waga jest proporcjonalna do częstości występowania terminu indeksującego w dokumencie i odwrotnie proporcjonalna do częstości dokumentowej (liczby dokumentów zawierających ten termin). Model przestrzeni wektorowej zaproponowany został w [Salton i in. 1975].

Realizacja zapytań polega na potraktowaniu zapytania jako pseudodokumentu i wyznaczeniu jego reprezentacji w tej samej przestrzeni, w której ulokowane zostały pozostałe elementy korpusu. Takie rozwiązanie pozwala na potraktowanie zarówno dokumentów, jak i samego zapytania jako punkty ulokowane w pewnej przestrzeni, co pozwala na wyznaczenie odległości pomiędzy zapytaniem i dokumentami i wybór tych, dla których tak określona miara przyjmuje wartość najmniejszą.

Przy stosowaniu przedstawionego podejścia metody uczenia maszynowego znajdują szczególne zastosowanie na etapie analizy zawartości informacyjnej dokumentów. Reprezentacja wyrazów w postaci punktów w wielowymiarowej przestrzeni, dokumentów zaś jako ich zbiorów pozwala na stosowanie wszystkich metod opartych na miarach odległości lub podobieństwa.

## 7. Modele probabilistyczne

Modele probabilistyczne zakładają sekwencyjny charakter tekstu. W zależności od celu badań tekst może być rozumiany jako ciąg liter, wyrazów czy sylab. Związki

między elementami sekwencji opisywane są za pomocą pojęć rachunku prawdopodobieństwa. Tego typu podejście w badaniach lingwistycznych zostało zapoczątkowane przez Andrieja Markowa, twórcę koncepcji procesów Markowa, w których rozważana jest sekwencja stanów pewnego systemu i prawdopodobieństwo wystąpienia stanu uzależnione jest od stanu bezpośrednio go poprzedzającego. W 1913 r. tego rodzaju model Markow wykorzystano do opisu sekwencji liter w tekstach rosyjskich [Manning, Schütze 1999].

Na bazie klasycznego modelu Markowa stworzona została koncepcja *ukrytego procesu Markowa* (HMM – *Hidden Markov Model*), w którym sekwencja kolejnych stanów systemu nie jest bezpośrednio obserwowalna, lecz dostępne są jedynie sekwencje wartości pewnej funkcji losowej przekształcającej ukryty dla obserwatora stan systemu w wartość obserwowalną.

Parametry klasycznego oraz ukrytego modelu Markowa szacowane są na podstawie zbioru uczącego. Ukryty model Markowa wykorzystywany jest z powodzeniem w zadaniach identyfikacji części mowy, gdzie ma szczególne znaczenie przy podejmowaniu decyzji dotyczących wyrazów wieloznacznych. Natomiast w językach azjatyckich modele tego typu stosowane są do wyznaczania granic słów.

## 8. Modele oparte na gramatykach formalnych

Gramatyka formalna jest opisem zasad budowy poprawnych wypowiedzi na bazie przyjętego słownika elementów składowych. Gramatyka formalna operuje dwiema grupami pojęć: pojęciami terminalnymi oraz pojęciami nieterminalnymi. Pojęcia terminalne to te, które mogą wystąpić w wypowiedzi. Pojęcia nieterminalne mają charakter abstrakcyjny, nie pojawiają się w wypowiedzi, lecz są zastępowane przez odpowiadające im symbole terminalne. Szczególne znaczenie ma symbol startowy, mający charakter nieterminalny i reprezentujący całą wypowiedź. Definicja gramatyki obejmuje również reguły pozwalające na sprawdzenie, czy rozpatrywany ciąg symboli terminalnych jest prawidłowy. Reguły te określane są mianem produkcji. Sprawdzając poprawność wypowiedzi, należy znaleźć ciąg produkcji pozwalających na jej wyprowadzenie z symbolu startowego. Identyfikacja produkcji pozwalających na wyprowadzenie analizowanego zdania pozwala również na dokonanie interpretacji jego elementów. Autorem koncepcji gramatyk formalnych jest Noam Chomsky. Wśród wielu publikacji prezentujących zagadnienia gramatyk formalnych i ich zastosowań warto wymienić prace: [Révész 1983; Clark i in. 2010].

Gramatyki formalne są narzędziem opisu i analizy zarówno języków sztucznych (np. języków programowania), jak i języków naturalnych. Szczególnie przydatną klasą gramatyk są gramatyki bezkontekstowe, w których produkcje mają postać:

symbol nieterminalny  $\rightarrow$  ciąg symboli terminalnych i/lub nieterminalnych,

czyli sposób traktowania symboli nieterminalnych nie zależy od ich kontekstu.



Zastosowanie reguł zdefiniowanych w ramach gramatyki bezkontekstowej pozwala na dokonanie rozbioru wypowiedzi na poszczególne elementy składowe. Sposób dokonywania rozbioru może zostać przedstawiony w postaci drzewa rozbioru.

O ile gramatyki bezkontekstowe dobrze radzą sobie z analizą języków sztucznych, o tyle języki naturalne, ze względu na ich nieregularność i dopuszczalność wielu alternatywnych sposobów konstrukcji wypowiedzi, stanowiły wyzwanie dla podejmowanych prób ich opisu. Koncepcją pozwalającą na zastosowanie produkcji gramatyk bezkontekstowych do opisu języków naturalnych są probabilistyczne gramatyki bezkontekstowe (określane również jako stochastyczne gramatyki bezkontekstowe). W rozwiązaniach tych z każdą produkcją powiązane jest prawdopodobieństwo jej wystąpienia. Prawdopodobieństwa te przypisywane są w taki sposób, aby suma prawdopodobieństw przypisanych do produkcji o tym samym poprzedniku równała się jedności. W gramatykach probabilistycznych istnieje wiele alternatywnych sposobów rozbioru analizowanej wypowiedzi. Jednakże biorąc pod uwagę prawdopodobieństwa produkcji, można wyznaczyć prawdopodobieństwo wystąpienia każdego drzewa rozbioru (i wybrać najbardziej prawdopodobny sposób interpretacji) i wybrać to, które jest najbardziej prawdopodobne. Budowa zestawu produkcji i przypisywanie im prawdopodobieństw może zostać przeprowadzone przez uczenie maszynowe, głównie przez zastosowanie algorytmów genetycznych.

Do znajdowania najbardziej prawdopodobnego drzewa rozkładu stosuje się zwykle algorytm inside-outside lub algorytm  $A^*$  [Manning, Schütze 1999]. Algorytm inside-outside pozwala na dokonanie wyboru zbioru produkcji maksymalizującego prawdopodobieństwo uzyskania zdań wchodzących w skład zbioru uczącego. Algorytm pracuje w trybie bez nauczyciela. Jego wadą jest skłonność do zatrzymywania się w minimach lokalnych optymalizowanej funkcji. Natomiast algorytm  $A^*$  należy do metod przeszukiwania grafów. W lingwistyce metoda ta jest stosowana do identyfikacji najbardziej prawdopodobnego drzewa rozkładu analizowanego zdania przy założeniu, że produkcje gramatyki typu PCFG są znane. Autorami koncepcji zastosowania algorytmu  $A^*$  do parsowania zdań w języku naturalnym są Klein oraz Manning [2003].

## 9. Modele w postaci sieci semantycznej

Sieć semantyczna w niniejszej pracy rozumiana będzie jako narzędzie opisu wyodrębnionego fragmentu rzeczywistości pozwalające na wyodrębnienie klas obiektów, definiowanie schematów ich opisów i przedstawienie relacji pomiędzy klasami oraz reprezentowanie rzeczywistych instancji obiektów i relacji. Definiowanie schematu opisu obiektu polega na przyjęciu zestawu własności charakteryzujących obiekt. Dla każdej własności może zostać ustalony typ przypisanych jej wartości oraz mogą zostać nałożone warunki precyzujące zakres dopuszczalnych wartości. Relacje opisywać mogą hierarchię klas oraz związki zachodzące pomiędzy klasami i obiektami.



Przyjęty za obowiązujący, akceptowany przez ogół użytkowników i pozwalający na jednoznaczny opis wymienianych informacji sposób opisu wyodrębnionej dziedziny określany jest terminem ontologii.

Zastosowanie ontologii znacznie podnosi jakość wyników eksploracyjnej analizy tekstów. Podejście wykorzystujące wiedzę dziedzinową w postaci ontologii nie wyklucza również stosowania metod uczenia maszynowego. Należy zauważyć, że algorytmy uczące się mogą być stosowane w trzech całkowicie różnych obszarach. Po pierwsze, mogą być stosowane na etapie konstrukcji ontologii. Zastosowanie metod eksploracyjnych pozwala na identyfikację klas oraz pozwala określić ich hierarchię. Analiza danych jest również przydatnym narzędziem identyfikacji związków pomiędzy klasami [Buitelaar i in. 2003]. Drugim zastosowaniem jest uczenie zasad transformacji tekstu do postaci sieci semantycznej. Tworzone na tym etapie rozwiązanie ma na celu analizę fragmentu tekstu i utworzenie instancji klasy, przypisanie właściwych wartości własnościom obiektu lub rozpoznanie relacji pomiędzy klasami. Trzecim zastosowaniem algorytmów uczenia maszynowego jest analiza danych tworzących sieć semantyczną.

Ontologiczny model zawartości informacyjnej dokumentu w sposób precyzyjny i jednoznaczny reprezentuje treść tekstu. Ważną zaletą tego podejścia jest możliwość wykorzystania modelu zarówno przez człowieka, jak i przez system komputerowy. Reprezentacja ontologiczna pozwala na realizację zapytań. Zapytania kierowane do sieci semantycznej formułowane są za pomocą przyjętej formalnej notacji. Przykładami tego typu rozwiązań jest język SPARQL (*Simple Protocol and RDF Query Language*) oraz OWL-QL (*Web Ontology Language – Query Language*).

## **10. Rola wiedzy dziedzinowej i uczenia maszynowego w procesie konstrukcji modeli zawartości informacyjnej**

Cechą wspólną wszystkich przedstawionych powyżej podejść jest możliwość budowy modelu na bazie danych w postaci korpusu dokumentów. Jednakże wszystkie przedstawione powyżej rozwiązania wymagają również dostarczenia wiedzy zewnętrznej (lingwistycznej lub dziedzinowej). Aspekt ten jest szczególnie istotny z punktu widzenia czasochłonności i kosztów budowy modeli. Zgromadzenie odpowiednich zasobów wiedzy oraz opracowanie właściwego sposobu jej reprezentacji jest szczególnie kosztownym elementem tworzonych systemów (w wymiarze zarówno czasowym, jak i finansowym). Z drugiej jednak strony próby szukania oszczędności na tym polu prowadzić mogą do istotnego spadku jakości konstruowanych rozwiązań.

Warto prześledzić rolę, jaką odgrywa wiedza zewnętrzna i uczenie maszynowe w poszczególnych klasach modeli. Zestawienie przedstawiano w tab. 1.

**Tabela 1.** Rola uczenia maszynowego i wiedzy zewnętrznej w modelach zawartości informacyjnej dokumentów tekstowych

Model	Uczenie maszynowe	Wiedza zewnętrzna
Model przestrzeni wektorowej	<ul style="list-style-type: none"> <li>• konstrukcja modelu na podstawie macierzy częstości</li> <li>• redukcja wymiaru modelu</li> <li>• analiza informacji zawartych w dokumentach i reprezentowanych przez punkty w przestrzeni</li> </ul>	<ul style="list-style-type: none"> <li>• listy terminów indeksujących</li> <li>• listy terminów nieistotnych (stop-lista)</li> <li>• synonimy</li> <li>• frazy</li> <li>• forma podstawowa wyrazów</li> </ul>
Model probabilistyczny	<ul style="list-style-type: none"> <li>• konstrukcja modelu</li> <li>• szacowanie prawdopodobieństwa wystąpienia wypowiedzi</li> </ul>	<ul style="list-style-type: none"> <li>• anotacja lingwistyczna wyrazów</li> </ul>
Model gramatyczny	<ul style="list-style-type: none"> <li>• identyfikacja reguł gramatycznych,</li> <li>• oszacowanie prawdopodobieństw dla poszczególnych reguł</li> <li>• parsowanie tekstu i wybór najbardziej prawdopodobnego drzewa rozbioru</li> </ul>	<ul style="list-style-type: none"> <li>• anotacja lingwistyczna i semantyczna wyrazów</li> <li>• definicja produkcji</li> </ul>
Model w postaci sieci semantycznej	<ul style="list-style-type: none"> <li>• wspomaganie procesu tworzenia ontologii</li> <li>• tworzenie reguł przekształcających fragmenty tekstu w elementy sieci semantycznej</li> <li>• analiza danych w postaci sieci semantycznej</li> </ul>	<ul style="list-style-type: none"> <li>• wiedza dziedzinowa w postaci ontologii</li> <li>• anotacja morfologiczna i semantyczna wyrazów</li> <li>• identyfikacja nazw własnych</li> <li>• reguły interpretacji fragmentów tekstu</li> </ul>

Źródło: opracowanie własne.

## 11. Lokalny i globalny charakter poszczególnych modeli zawartości informacyjnej

Porównując zawartość informacyjną korpusu dokumentów oraz zasięg obszaru opisywanego przez model zawartości informacyjnej, za celowe należy uznać wyróżnienie modeli o charakterze:

- globalnym – w których zakres obszaru opisywanego przez model pokrywa się z zakresem całego dokumentu,
- lokalnym – które opisują jedynie zawartość pewnego, zwykle niewielkiego, fragmentu dokumentu.

Wśród przedstawionych powyżej koncepcji modele oparte na koncepcji przestrzeni wektorowej oraz wykorzystujące sieci semantyczne mają charakter globalny. Natomiast modele probabilistyczne oraz wykorzystujące gramatyki formalne przydatne są raczej do opisu fragmentu dokumentu – można je więc określić jako lokalne.

Lokalny bądź globalny charakter modelu nie przesądza o ocenie danego rozwiązania. Dla każdego rodzaju modelu przewidziany jest odmienny zakres zadań.

Sz szczególnie interesujące jest porównanie dwóch podejść pozwalających na reprezentację całości informacji zawartych w korpusie. Model oparty na przestrzeni wektorowej oraz sieć semantyczna różnią się praktycznie wszystkimi cechami: sposobem reprezentacji informacji, czasem i kosztem budowy czy sposobem analizy reprezentowanych informacji. Z tego powodu dokonanie właściwego wyboru jest w tym przypadku szczególnie ważne.

Podejścia stosowane w modelach lokalnych są znacznie bardziej do siebie zbliżone pod względem możliwości i kosztów konstrukcji.

Należy również zwrócić uwagę na możliwości łącznego wykorzystania modeli. Szczególnie przydatne może być wykorzystanie modeli lokalnych przy konstrukcji modeli globalnych. Podejście takie może być stosowane w szerokim zakresie przy budowie modeli w postaci sieci semantycznej. Natomiast przydatność modeli lokalnych w trakcie tworzenia modelu opartego na przestrzeni wektorowej jest ograniczona (do identyfikacji fraz i przekształcenia słów do postaci podstawowej).

## 12. Podsumowanie

Głównym celem niniejszej pracy było przedstawienie roli uczenia maszynowego w procesie pozyskiwania informacji z dokumentów tekstowych. Założono, że zadanie to realizowane jest za pośrednictwem modelu zawartości informacyjnej dokumentu. Przedstawione charakterystyki poszczególnych rozwiązań wskazują na bardzo duże zróżnicowanie poszczególnych koncepcji.

W przypadku podejść opartych na koncepcji przestrzeni wektorowej uczenie maszynowe odgrywa szczególnie istotną rolę na etapie zastosowania modelu. W modelach probabilistycznych oraz opartych na gramatykach formalnych i na sieciach semantycznych uczenie maszynowe jest bardzo przydatne już na etapie ich tworzenia. Pozyskiwanie informacji z dokumentów tekstowych pozwala na wykorzystanie zarówno nadzorowanych (uczenie z nauczycielem), jak i nienadzorowanych (uczenie bez nauczyciela) metod uczenia maszynowego.

Wydaje się, że podstawową przesłanką przemawiającą za korzystaniem z uczenia maszynowego w zadaniach pozyskiwania informacji z zasobów tekstowych jest brak możliwości formalnego ujęcia wszystkich poprawnych wypowiedzi sformułowanych w języku naturalnym za pomocą precyzyjnych reguł. Ze względu na złożony charakter wypowiedzi oraz dużą ilość zasobów tekstowych rozwój zastosowań metod uczenia maszynowego wymaga rozwiązań charakteryzujących się dużą mocą obliczeniową i dużymi możliwościami w zakresie przechowywania danych.

## Literatura

Buitelaar P., Cimiano Ph., Magnini B., *Ontology Learning from Text: An Overview*, 2003, <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.70.3041>.

- Clark A., Fox C., Lappin S. (red.), *The Handbook of Computational Linguistics and Natural Language Processing*, Blackwell Publishing Ltd, 2010.
- Farkas R., *Machine learning techniques for applied information extraction*, Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and the University of Szeged, June 2009, [https://docs.google.com/viewer?url=http%3A%2F%2Fwww.sci.u-szeged.hu%2Ffokozatok%2FPDF%2FFarkas\\_Richard%2Fthesis.pdf](https://docs.google.com/viewer?url=http%3A%2F%2Fwww.sci.u-szeged.hu%2Ffokozatok%2FPDF%2FFarkas_Richard%2Fthesis.pdf).
- Klein D., Manning C., *A\* parsing: fast exact viterbi parse selection*, Proceedings of the Human Language Technology Conference and the North American Association for Computational Linguistics (HLT-NAACL 2003), Main Papers, Edmonton, May-June 2003.
- Manning C., Schütze H., *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA: May 1999.
- Révész G., *Introduction to Formal Languages*, McGraw-Hill Book Company, 1983.
- Salton G., Wong A., Yang C.S., *A vector space model for automatic indexing*, „Communications of the ACM” 1975, vol. 18, no 11.

## LEARNING-BASED SYSTEMS OF INFORMATION EXTRACTION FROM TEXTUAL RESOURCES

**Summary:** The main aim of this work is the presentation, classification and evaluation learning-based systems of information extraction from textual resources. In the initial part of the paper the concept of the learning-based system and the problem of information extraction are presented. The next part of the article presents the structure and the functioning of the information extraction solutions. The model of information content is a key element of such systems. Its characteristics and types are the principal subject of the next point. The following point presents the role of external knowledge and machine learning approach in various solutions. In the next part of the article some remarks concerning local or global character of the individual solutions are presented.

**Keywords:** information extraction from textual resources, text mining.