Załącznik 2a

Piotr Gawrysiak
Instytut Informatyki
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

Warszawa, 1 marca 2015

AUTOREFERAT

1. Imię i nazwisko: Piotr Tadeusz Gawrysiak

2. Posiadane dyplomy, stopnie naukowe/artystyczne

1998 tytuł magistra inżyniera informatyki w zakresie budowy i oprogramowania maszyn cyfrowych uzyskany na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej w Warszawie. Tytuł pracy magisterskiej „*Przeszukiwanie tekstowych baz danych w języku polskim*". Promotor: prof. Henryk Rybiński.

2001 tytuł magistra zarządzania i administracji w zakresie zarządzania gospodarczego uzyskany na Wydziale Zarządzania Uniwersytetu Warszawskiego w Warszawie. Tytuł pracy magisterskiej „*Data mining. Automatyczna eksploracja baz danych*". Promotor: prof. Witold Chmielarz.

2002 stopień naukowy doktora nauk technicznych uzyskany na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej w Warszawie. Tytuł rozprawy doktorskiej „*Automatyczna kategoryzacja dokumentów*". Promotor: prof. Henryk Rybiński.

2010 stopień naukowy doktora habilitowanego nauk humanistycznych w zakresie bibliologii uzyskany na Wydziale Historycznym Uniwersytetu Warszawskiego w Warszawie. Tytuł rozprawy habilitacyjnej „*Cyfrowa rewolucja. Rozwój cywilizacji informacyjnej*".

3. Informacje o dotychczasowym zatrudnieniu w jednostkach naukowych/artystycznych

1999-2002	Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Instytut Informatyki, doktorant
2001-2002	Ulster University, Northern Ireland Knowledge Engineering Laboratory, visiting researcher
2002-2011	Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Instytut Informatyki, adiunkt
2003-2004	Wyższa Szkoła Informatyki i Ekonomii TWP w Olsztynie, adiunkt
2004-2006	Politechnika Warszawska, Laboratorium Technologii Mobilnych BRAMA, kierownik
2007-2010	Uniwersytet Warszawski, Wydział Historyczny, wykłady zaproszone (4 semestry)

2008-2012 Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Instytut Informatyki, Zastępca Dyrektora ds. Naukowych

2011-obecnie Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Instytut Informatyki, profesor nadzwyczajny

4. Wskazanie osiągnięcia wynikającego z art. 16 ust. 2 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. Nr 65, poz 595 ze zm.)

4.1 Tytuł osiągnięcia naukowego:

"Metody semantycznego przetwarzania i wizualizacji informacji"

4.2 Osiągnięcie naukowe – złożone z jednotematycznego cyklu publikacji naukowych oraz oryginalnego osiągnięcia technicznego

a) Wykaz publikacji wchodzących w skład osiągnięcia naukowego

[P1] Gawrysiak Piotr, Gancarz Łukasz, Okoniewski Michał: Recording word position information for improved document categorization, in: Proceedings of the 6th Pacific and Asia Conference on Knowledge Discovery and Data Mining PAKDD 2002, Workshop on Text Mining, Taipei, Tajwan, 2002, s. 25-34

[P2] Gawrysiak Piotr, Rybiński Henryk, Gajda Damian, Gołębski Marcin: Extending open source software solutions for CRM text mining, in: Proceedings of the IADIS International Conference WWW/Internet ICWI 2004, vol. III, Madryt, Hiszpania, 2004, s. 869-872

[P3] Gawrysiak Piotr: Extending traditional Wiki systems with geographical content, in: Proceedings of 6th IFIP Conference on e-Commerce, e-Business and e-Government, I3E 2006, Suomi R. *[et al.]* (*eds*), Turku, Finlandia, Springer, 2006, s. 292-302

[P4] Gawryjołek Jakub, Gawrysiak Piotr: The analysis and visualization of entries in Wiki services, in: Proceedings of 5th International Atlantic Web Intelligence Conference AWIC 2007, Advances in Intelligent Web Mastering / Węgrzyn-Wolska Katarzyna M. *[et al.]* (*eds.*), Advances in Soft Computing, Fontaineblau, Francja, Springer, 2007, s. 118-123

[P5] Gawrysiak Piotr, Rybiński Henryk, Skonieczny Łukasz, Więch Przemysław Bartłomiej: AMI-SME: An exploratory approach to knowledge retrieval for SMEs, in: Proceedings of the Third International Conference on Autonomic and Autonomous Systems ICAS 2007, Athens, Grecja, 2007, s.1-6

[P6] Gawrysiak Piotr, Protaziuk Grzegorz Michał, Rybiński Henryk, Delteil Alexander: Text onto miner - a semi automated ontology building system, in: Proceedings of 17th

International Symposium on Methodologies for Intelligent Systems ISMIS 2008, An. A.*[et al.]* (*eds.*) Lecture Notes in Artificial Intelligence, vol. 4994, Toronto, Kanada, Springer, 2008, s. 563-573

[P7] Gawrysiak Piotr, Ryżko Dominik Paweł: Analysing and Visualizing Polish Scientific Community, in: Business and Engineering Applications of Intelligent and Information Systems / Setlak G., Markov K. (*eds.*), ITHEA, 2011, s. 47-56

[P8] Kołaczkowski Piotr, Gawrysiak Piotr: Extracting Product Descriptions From Polish E-Commerce Websites Using Classification and Clustering, in: Proceedings of 19th International Symposium on Methodologies for Intelligent Systems ISMIS 2011, Foundations of Intelligent Systems / Kryszkiewicz Marzena *[et al.]* (*eds.*), Lecture Notes in Artificial Intelligence, vol. 6804, Warszawa, Springer, 2011, s. 456-464

[P9] Burdziej Jan, Gawrysiak Piotr: Using Web Mining for Discovering Spatial Patterns and Hot Spots for Spatial Generalization, in: Proceedings of 20th International Symposium on Methodologies for Intelligent Systems ISMIS 2012, Foundations of Intelligent Systems / Chen Li *[et al.]* (*eds.*), Lecture Notes in Artificial Intelligence, vol. 7661, Macau, Chiny, Springer, 2012, s. 172-181

[P10] Gawrysiak Piotr, Ryżko Dominik Paweł, Więch Przemysław B., Kozłowski Marek: Retrieval and Management of Scientific Information from Heterogeneous Sources , in: Intelligent Tools for Building a Scientific Information Platform / Bembenik Robert *[et al.]* (*eds.*), Studies in Computational Intelligence, vol. 390, Springer, 2012, s. 37-48

[P11] Twardowski Bartłomiej, Gawrysiak Piotr: Domain Dependent Product Feature and Opinion Extraction Based on E-Commerce Websites,, in: Multimedia and Internet Systems: Theory and Practice / Zgrzywa Aleksander, Choroś Kazimierz, Siemiński Andrzej (*eds.*), Advances in Intelligent Systems and Computing, vol. 183, Springer, 2013 s. 261-270

b) oryginalne osiągnięcia techniczne wchodzące w skład osiągnięcia naukowego

[T1] AMI-SME, pakiet oprogramowania zaimplementowany w ramach projektu *Analysis of marketing information for small- and medium-sized enterprises*, projekt badawczy CRAFT no. 005875 w ramach programu FP6, Instytut Informatyki Politechniki Warszawskiej, Warszawa, 2004-2007 *(bez interfejsu użytkownika)*

[T2] *Text-Onto-Miner*, pakiet oprogramowania zaimplementowany w ramach projektu badawczego dla France Telecom n. 46 132 897, Instytut Informatyki Politechniki Warszawskiej, Warszawa, 2006-2008

4.3 Omówienie publikacji oraz osiągnięć technicznych

Badania nad sztuczną inteligencją należą do szczególnie istotnych obszarów współczesnej nauki i są obecnie intensywnie rozwijane. Stworzenie urządzenia, czy też programu komputerowego który można by nazwać „inteligentnym", często jest uznawane za swego rodzaju finalny cel informatyki. Stanowi to problem tym bardziej interesujący, ponieważ nie brak badaczy argumentujących, iż jest

to cel wręcz niemożliwy do osiągnięcia (patrz np. polemika pomiędzy Rogerem Penrose [1] i Johnem McCarthy [2]). Inni wyrażają z kolei pogląd, iż tego rodzaju badania powinny być objęte moratorium ze względu na zagrożenia, które zaawansowana sztuczna inteligencja może stworzyć dla ludzkiej cywilizacji.

Nie ulega wątpliwości, iż jednym z etapów, prowadzących do stworzenia systemów sztucznej inteligencji, będzie opanowanie metod automatycznego pozyskiwania, przetwarzania i reprezentacji wiedzy. Na pozór problem przetwarzania wiedzy został obecnie już dość dobrze opanowany. Pomimo jednak, iż nauka dysponuje dość bogatym instrumentarium, które można określić mianem syntaktycznego, to większość dostępnej ludziom wiedzy pozostaje wciąż zapisana w formie nieuporządkowanej. W pewnych przypadkach możliwe jest zapisywanie wiedzy w sposób formalny, a tak reprezentowana może być analizowana semantycznie (co obejmuje choćby wnioskowanie), jednak większość zasobów wiedzy to dokumenty w języku naturalnym i w formie graficznej. Wydobycie struktury i znaczenia „ukrytego" w tych źródłach, niezbędne do wyposażenia systemów sztucznej inteligencji we własną wiedzę, jest głównym przedmiotem badań naukowych habilitanta.

W pracach habilitanta stanowiących osiągnięcie naukowe wyróżnić można cztery ściśle powiązane ze sobą grupy tematyczne : pozyskiwanie informacji semantycznej ze źródeł pełnotekstowych, analiza źródeł pełnotekstowych w sieci WWW, półautomatyczne budowanie sieci semantycznych, oraz przetwarzanie i wizualizacja informacji heterogenicznych. Zostaną one pokrótce omówione poniżej.

a) Pozyskiwanie informacji semantycznej ze źródeł pełnotekstowych

Wykorzystanie narzędzi i algorytmów automatycznej analizy języka naturalnego (ang. *NLP – natural language processing*) do analizy dużych baz nieustrukturalizowanych danych stanowi obecnie jedno z podstawowych metod eksploracji danych. Proces ten bywa określany niekiedy mianem *text mining*, dla odróżnienia od "klasycznej" eksploracji danych (ang. *data mining*), której przedmiotem zainteresowania są przede wszystkim dane ustrukturalizowane, w szczególności zaś dane nietekstowe pochodzące bezpośrednio z tzw. systemów transakcyjnych (takich jak systemy sprzedaży detalicznej, systemy bilingowe czy też systemy bankowości elektronicznej).

Jedne z pierwszych eksperymentów habilitanta związanych z eksploracją danych o innym, niż ustrukturalizowany, charakterze związane są z działalnością firm telekomunikacyjnych. Pierwsze projekty badawcze, w których podejmowano próby zastosowania metod eksploracji danych w środowisku telekomunikacyjnym realizowane były, zarówno w Polsce jak i za granicą, na początku ubiegłej dekady. W szczególności wymienić tu należy projekty prowadzone we współpracy pomiędzy Polską Telefonią Cyfrową sp. z o.o. (obecnie T-Mobile Polska), a Politechniką Warszawską, w których brał udział habilitant. W projektach tych, co stanowiło w owym czasie znaczącą nowość, przedmiotem analizy były dane generowane bezpośrednio przez infrastrukturę sieci telekomunikacyjnej, takie jak dane z systemu sygnalizacji SS7 czy też statystyki obciążenia elementów sieci. Prace badawcze zakończone zostały z sukcesem, część rozwiązań została wdrożona przez Polską Telefonię Cyfrową, zaś metoda predykcji obciążenia komórek sieci,

współautorstwa habilitanta, została ponadto opatentowana[1]. Dzięki temu zespół Instytutu Informatyki Politechniki Warszawskiej zyskał doświadczenie, które pozwoliło na efektywne poszukiwanie innych projektów badawczych zlecanych przez firmy telekomunikacyjne. W szczególności nawiązano współpracę z firmą Polkomtel S.A., gdzie przedmiotem badań było wspieranie pracy biura obsługi klienta operatora telekomunikacyjnego, w ramach procesu budowania relacji z klientem (ang. C*RM – customer relationship management).* W kierowanym przez habilitanta projekcie badawczym zbudowano modułowe narzędzie pozwalające na automatyczną analizę dokumentacji przetwarzanej przez *call center*, w tym w szczególności klasyfikację korespondencji z klientem w celu oceny stopnia satysfakcji z usług operatora oraz stworzenie innych mechanizmów automatycznych – takich jak m.in. odpowiadanie na typowe pytania użytkowników bezpośrednio przez system, bez udziału pracowników. Istotnym problemem, jaki wymagał rozwiązania, był charakter dokumentów, które podlegały przetwarzaniu przez system. Nie poddawały się one prostej analizie przy wykorzystaniu dostępnych w owym czasie algorytmów przetwarzania języka naturalnego takich jak klasyfikatory i algorytmy grupowania wykorzystujące model frekwencyjny reprezentacji tekstu [3], [4], zawierały bowiem niezwykle wysoką liczbę błędów, zarówno ortograficznych (brakujące litery, przeinaczenia słów, specyficzne skróty i in.), jak i gramatycznych, czy też semantycznych, co wynikało ze specyfiki pracy wpisujących je pracowników *call-center*. Ponadto zawierały tekst w języku polskim, który nie był wówczas dobrze wspierany przez dostępne pakiety wolnego oprogramowania[2] NLP, takie jak m.in. *stemmery*, w tym powszechnie używany w odniesieniu do tekstów w języku angielskim *stemmer* Portera [5]. Warto zaznaczyć, iż stosowane współcześnie narzędzia analizy języka polskiego, w tym te opracowywane w Instytucie Podstaw Informatyki PAN, były w tym czasie dopiero tworzone – patrz np. [6] i [7]. W ramach projektu zaimplementowano algorytmy wstępnej obróbki dokumentów, wykorzystujące m.in. zaproponowaną przez habilitanta nową metodę reprezentacji tekstu (opisaną w artykule [P1]). W reprezentacji tej, nie jest całkowicie tracona informacja o położeniu słów w obrębie analizowanego dokumentu. Dla danego dokumentu $D=(w_1, w_2, ..., z_1, ..., w_n, z_m)$ wspomniana reprezentacja jest parą $(\boldsymbol{F}, \boldsymbol{S})$ gdzie $\boldsymbol{S}$ to wektor skalujący o wartościach odpowiadających modelowi frekwencyjnemu, zaś $\boldsymbol{F}$ to zbiór funkcji gęstości f_{Vi} takich że $f_{v_i}(k)=\sum_{j=k-r}^{k+r}(1\ jeżeli\ w_j=v_i, v_i\in V; 0\ wpw)/\alpha_i$ oraz $\sum_{j=k-r}^{n} f_{v_i}=1$ Funkcje te mogą być następnie porównywane bezpośrednio, m.in. za pomocą dywergencji Kullbacka-Leiblera, lub też po aproksymacji histogramami. Opracowano także dostosowane do specyfiki dokumentów algorytmy korekcji błędów ortograficznych, czy też narzędzia wizualizacji tematyki kolekcji dokumentów tekstowych, wykorzystujące krzywe Zipfa. Rozszerzono także funkcjonalność pakietu Apache Lucene, który zastosowany został jako podstawowy system indeksowania dokumentów w projekcie. Artykuł [P2] zawiera opis wspomnianego projektu, zaś oprogramowanie stworzone na jego potrzeby zostało wykorzystane przy realizacji kolejnych projektów badawczych w Instytucie Informatyki PW (m.in. realizowanych we współpracy z Organizacją Narodów Zjednoczonych do

[1] Piotr Gawrysiak, Michał Okoniewski, Sposób estymowania ruchu w komórce telefonii komórkowej, patent nr 198310, Biuletyn Urzędu Patentowego nr 09/2002

[2] Wykorzystanie zaś i rozwój, jeszcze wówczas niezbyt popularnego w środowisku przemysłowym, oprogramowania open source było jednym z założeń projektu

spraw Wyżywienia i Rolnictwa (FAO)[3], oraz stało się podstawą do rozwoju systemów opisanych w następnym punkcie.

b) Analiza źródeł pełnotekstowych w sieci WWW

Pierwsza połowa ubiegłej dekady to okres, w którym sieć Internet poczyna być postrzegana nie tylko jako medium służące bezpośredniej komunikacji, ale także jako swego rodzaju globalna baza wiedzy. Spowodowało to wzrost zainteresowania pozyskiwaniem i obróbką informacji pochodzących bezpośrednio z zasobów sieci WWW. Internet stanowi źródło danych nieporównanie trudniejsze w analizie automatycznej w stosunku do tych utrzymywanych lokalnie, takich jak korporacyjne bazy danych. Wynika to zarówno z jego rozmiaru, jak i na stopnia nieuporządkowania i zmienności. Jednocześnie jest on jednak zasobem niezwykle atrakcyjnym, poprzez swoją dostępność i kompletność.

W badaniach nad automatycznym wykorzystaniem informacji zgromadzonych w globalnej sieci wyróżnić można dwa trendy. Pierwszy z nich obejmuje inicjatywy zmierzające do uporządkowania zasobów Internetu i tym samym stworzenie globalnej pajęczyny semantycznej (ang. *semantic web*) [8], która w naturalny sposób miałaby być wykorzystywana przez systemy automatyczne. Pomimo wielu wysiłków dokonywanych na przestrzeni ostatnich kilkunastu lat nie doprowadziło to jak na razie do powstania zbyt wielu użytecznych źródeł danych. Jednym z powodów jest tu konieczność bezpośredniego zaangażowania twórców treści publikowanej w sieci do jej strukturalizowania i uzupełniania dodatkową informacją semantyczną, której użyteczność nie jest widoczna *ad hoc*.

Drugie podejście obejmuje konstruowanie systemów wykorzystujących mechanizmy sztucznej inteligencji, w tym eksploracji danych, do analizy istniejących, nieustrukturalizowanych zasobów sieci WWW. W trend ten wpisują się badania nad tworzeniem efektywnych systemów wyszukiwawczych dostosowanych do specyfiki zasobów sieciowych. Ich wynikiem są, poczynając już od pierwszych prac Brina i Page'a [9], komercyjne wyszukiwarki sieciowe, powszechnie dziś używane przez wszystkich użytkowników sieci Internet. Jednocześnie prowadzono także badania nad narzędziami wspierającymi proces przeglądania i monitorowania zasobów sieciowych przez użytkowników, co można określić mianem procesu wyszukiwania eksploracyjnego (ang. *exploratory search*) [10], stosowane jest tu także określenie *web mining*. Odmiana tego procesu dostosowana zaś do wymogów środowiska biznesowego, zdefiniowana została przez Hackathorna w [11] i określona jako *web farming*.

Powyższe podejście zostało wykorzystane w projekcie badawczym AMI-SME, realizowanym w ramach szóstego programu ramowego Komisji Europejskiej m.in. w Instytucie Informatyki Politechniki Warszawskiej. Celem projektu było opracowanie narzędzi (zarówno procedur i rozwiązań organizacyjnych, jak i oprogramowania) wspierających małe i średnie przedsiębiorstwa w procesie internacjonalizacji, to jest rozszerzania zakresu działalności na rynki międzynarodowe. Główną barierą w takim procesie, jaką musi pokonać przedsiębiorstwo, jest wykonanie analizy rynków docelowych (tj. identyfikacja potencjału rynku, produktów i przedsiębiorstw konkurencyjnych, grup odbiorców itp.), co nierzadko wymaga zaangażowania środków

[3] Wyniki wybranych wspólnych projektów prezentowano na konferencji European Conference on Digital Libraries w Bath w 2004 roku

przekraczających możliwości niewielkich firm. Jeszcze kilkanaście lat temu tego rodzaju analiza byłaby właściwie niemożliwa bez udziału lokalnych podmiotów na każdym z docelowych rynków, których zadaniem byłaby akwizycja i wstępna obróbka informacji marketingowej. Metodyka opracowana w ramach projektu AMI-SME pozwala na wykonanie tego zadania dzięki wykorzystaniu zasobów informacyjnych Internetu, sam zaś proces zbierania, monitorowania i uaktualniania oraz organizacji pozyskanej informacji możliwy jest dzięki narzędziu AMI-SME [T1], jakie zostało zaimplementowane przez kierowany przez habilitanta zespół badawczy[4]. Narzędzie to łączy funkcjonalność klasycznej wyszukiwarki internetowej z narzędziami automatycznej analizy tekstu, pozwalając na szybkie, półautomatyczne zbudowanie ontologii produktowej, która następnie może być wykorzystana do wspierania procesu wyszukiwania informacji. Co istotne, w odróżnieniu od innych dostępnych wówczas systemów takich jak np. Protégé [12], stworzone narzędzie bazuje m.in. na bezpośredniej analizie wyników wyszukiwania, rozszerzając model zaproponowany przez Maedche [13], i tym samym nie wymagając od użytkownika specjalistycznej wiedzy, związanej z problematyką budowania sieci semantycznych. Implementacja systemu AMI-SME wymagała między innymi stworzenia dedykowanej, elastycznej bazy dokumentów tekstowych oraz opracowania algorytmów eksploracji danych tekstowych w środowisku wielojęzycznym (m.in. zaproponowane przez habilitanta algorytmy streszczania, grupowania i automatycznego etykietowania kolekcji dokumentów). W algorytmach tych dokumenty tekstowe transformowane są do postaci grafu skierowanego, którego węzły odpowiadają jednostkom funkcjonalnym tekstu (np. zdaniom lub akapitom). Następnie algorytmy analizy sieciowej, takie jak algorytm PageRank, czy też metoda HITS Kleinberga [14] mogą być zastosowane do wskazania najbardziej istotnych elementów grafu i tym samym najbardziej reprezentatywnych fragmentów dokumentu. Poza samą funkcjonalnością wyszukiwania informacji *ad hoc*, opartą przede wszystkim o wykorzystanie API serwisów Google Search i Yahoo Search, narzędzie AMI-SME wyposażono także we własny moduł akwizycji (ang. *harvesting*) umożliwiający automatyczne zbieranie informacji z dowolnych zasobów sieciowych, w tym także nie zaindeksowanych przez komercyjne serwisy wyszukiwawcze. Ponadto narzędzie to pozwala na śledzenie zmian w wynikach działania tychże serwisów, wynikających z modyfikacji wprowadzanych w źródłowych zasobach sieciowych, zgodnie z założeniami modelu *web mining*, co w tym czasie stanowiło istotną nowość. Wreszcie, część funkcjonalności AMI-SME dostępna jest bezpośrednio z poziomu interfejsu użytkownika przeglądarki internetowej, dzięki zaimplementowanemu w ramach projektu rozszerzeniu (ang. *plugin*) przeglądarki Mozilla Firefox. Wyniki prac zostały opisane m.in. w artykule [P5][5], zaś samo narzędzie wdrażane było we Włoszech, Austrii i Niemczech przez firmy uczestniczące w konsorcjum projektu.

c) Półautomatyczne budowanie sieci semantycznych

Opisane powyżej prace badawcze dotyczą głównie problematyki akwizycji informacji ze źródeł pełnotekstowych. Pokrewnym zagadnieniem, dotyczącym kolejnego etapu procesu wykorzystania tych informacji w praktyce, w szczególności w środowisku przemysłowym, jest strukturalizacja

[4] Zespół kierowany przez habilitanta implementował części analityczne i tzw. “logikę biznesową” systemu, interfejs użytkownika był w większości opracowywany przez innych członków konsorcjum projektu.

[5] Artykuł ten, prezentowany na międzynarodowej konferencji „ICAS 2007 - International Conference on Autonomic and Autonomous Systems” w Grecji otrzymał wyróżnienie „Best Paper Award”.

pozyskanej informacji w postaci sieci semantycznej (*ontologii*) oraz późniejsza „pielęgnacja” takiej sieci. Jak to opisano w [P6], jest to zagadnienie tym bardziej istotne, iż o ile rośnie zapotrzebowanie na ontologie (co związane jest ze wzrostem liczby usług w sieci Internet, wymagających wysokiej jakości metadanych semantycznych), to samo tworzenie sieci semantycznych pozostaje procesem żmudnym i kosztownym. Co więcej, większość istniejących w momencie opublikowania wspomnianego artykułu narzędzi automatyzujących ten proces, to metody i algorytmy specjalizowane, które możliwe są do wykorzystania jedynie w określonej dziedzinie wiedzy, czy też w odniesieniu tylko do wybranych rodzajów źródeł informacji. Narzędzia takie jak np. [15] i [16] można określić mianem „knowledge-rich”, jako że ich przystosowanie do nowej domeny wiedzy wymaga znaczącego wysiłku. Jak już wspomniano narzędzie AMI-SME, wspomniane w poprzednim punkcie, zawierało pewne proste mechanizmy wspierające budowanie sieci semantycznej wykorzystujące metody eksploracji danych, które nie wymagają tego rodzaju wstępnego przygotowania, i tym samym określane mogą być jako „knowledge-poor”. Doświadczenie zdobyte przy jego implementacji pozwoliło habilitantowi na realizację kolejnego projektu badawczego, tym razem już bezpośrednio dotyczącego ontologii, który został zlecony Instytutowi Informatyki Politechniki Warszawskiej przez francuskiego operatora telekomunikacyjnego France Telecom. W ramach tego projektu habilitant opracował wraz z zespołem metodę półautomatycznego budowania ontologii z dowolnych źródeł pełnotekstowych oraz wspierający ją pakiet oprogramowania TOM – Text-Onto-Miner [T2].

Metoda ta pozwala na szybkie i automatyczne tworzenie roboczych wersji ontologii dzięki analizie baz danych dokumentów pełnotekstowych (takich jak bazy dokumentacji technicznej; przy czym nie jest tu czynione żadne wstępne założenie o dziedzinie semantycznej tekstu), oraz zapewnia narzędzia ułatwiające późniejszą obróbkę powstającej sieci semantycznej. Habilitant opracował nowe algorytmy wstępnego przetwarzania dokumentów tekstowych, ekstrakcji pojęć i słów kluczowych, wreszcie zaś automatycznego odnajdowania i określania charakteru związków semantycznych pomiędzy pojęciami oraz wizualizacji sieci takich związków, w tym identyfikacji homonimów i synonimów, oraz automatycznego budowania taksonomii. W algorytmach tych dokumenty tekstowe transformowane są do postaci zbioru transakcji, zawierających słowa w nich występujące. W zależności od pożądanego poziomu granularności, za transakcję uznawany jest pojedynczy akapit dokumentu lub też zdanie, ew. inna jednostka struktury tekstu. Dzięki temu możliwe jest zastosowanie wysoko wydajnych algorytmów odkrywania zbiorów częstych do analizy współwystępowania słów. W szczególności możliwe jest m.in. określenie słów występujących w podobnych kontekstach, będących tym samym synonimami, lub też identyfikacja różnic pomiędzy stosunkowo podobnymi zbiorami słów w celu odkrycia potencjalnych homonimów. Algorytmy te zostały następnie zaimplementowane jako moduły elastycznego systemu analizy tekstu. W systemie tym możliwe jest definiowanie potoków przetwarzania danych przy wykorzystaniu graficznego interfejsu użytkownika, pozwalającego na wygodną analizę kolekcji dokumentów osobom przyzwyczajonym do komercyjnych pakietów eksploracji danych takich jak SAS i Statistica. Wspomniany artykuł [P6] zawiera opis systemu TOM wraz z przykładami jego zastosowania.

System ten został następnie wykorzystany w innych projektach badawczych prowadzonych w Instytucie Informatyki, w tym w szczególności podczas realizacji strategicznego projektu SYNAT

NCBiR[6], w którym moduły analizy języka naturalnego zostały użyte do zbudowania platformy repozytoryjnej (obejmującej m.in. bazę publikacji naukowych wraz z narzędziami analizy bibliometrycznej – patrz opis w artykule [P10]) $\Omega\Psi^{R}$.

W ostatnim okresie habilitant kontynuował badania nad automatycznym organizowaniem informacji semantycznej pochodzącej ze źródeł pełnotekstowych i sieciowych, koncentrując się na zastosowaniach w handlu elektronicznym (ang. *e-commerce*), zyskującym w ostatnich latach bardzo duże znaczenie gospodarcze, wraz z rosnącą rolą transakcji elektronicznych zarówno na rynku detalicznym, jak i B2B (*business-to-business*). Upowszechnienie „sklepów internetowych", oraz serwisów aukcyjnych spowodowało udostępnienie w sieci Internet dużych ilości informacji dotyczących cech produktów przemysłowych. Są to jednak informacje nieustrukturalizowane, toteż ich wykorzystanie w systemach automatycznych – na przykład służących porównywaniu cech użytkowych produktów czy też pozwalających na śledzenie zmian cen i popularności towarów – natrafia na problemy, o podobnym charakterze jak te, które były przedmiotem analizy w opisanych powyżej pracach habilitanta. Ostatnie eksperymenty dotyczą tu m.in. automatycznej ekstrakcji i obróbki cech produktów z ich opisów w systemach handlu elektronicznego oraz analizy opinii użytkowników dotyczących tych cech. W tym celu niezbędne jest automatyczne pozyskiwanie opisów produktów ze stron WWW platform handlu elektronicznego. Z racji wielości istniejących źródeł i ich zróżnicowania, jest to zadanie w którym nie mają zastosowania algorytmy oparte o języki specyfikacji wzorców, zwykle stosowane do ekstrakcji tekstu z zasobów sieciowych, takie jak np. [17] i [18]. Jednocześnie jednak podejścia wykorzystujące metody uczenia maszynowego takie jak [19], [20] i [21] okazują się także nieefektywne, z racji wielości sposobów prezentacji graficznej zasobów sklepów internetowych, czy też konieczności ręcznego etykietowania stron należących do zbioru trenującego. Niezbędne było zatem stworzenie autorskiego,prototypowego systemu ekstrakcji opisów produktów, opisanego w artykule [P8]. W systemie tym, po wykonaniu wstępnej filtracji zawartości tekstowej analizowanej strony WWW, budowane jest jej drzewo DOM (*ang. document object model*). Następnie, poczynając od liści, węzły drzewa są ze sobą łączone, tak aby uzyskać fragmenty tekstu o zadanej minimalnej długości. Fragmenty te są następnie klasyfikowane. Aby uprościć dostosowanie systemu do analizy arbitralnie wybranych kategorii produktów, zastosowano autorski algorytm klasyfikacji w którym docelowe kategorie definiowane są przez zbiory ważonych słów kluczowych. Prawdopodobieństwo *p* tego, iż analizowany fragment tekstu *f* należy do kategorii opisanej przez zbiór słów kluczowych $K=\{k_1...k_n\}$ i ich wag $W=\{w_1...w_n\}$ jest szacowane następująco: $p(f,K,W)=(\sum_{i=1}^{n} sgn\, w_i |w_i n(k_i,f)|^{\alpha})/N(f)$ gdzie *n(k,f)* to liczba wystąpień słowa *k* we fragmencie *f*, *N(f)* określa liczbę słów we fragmencie f, zaś $\alpha>0$ pozwala modyfikować wrażliwość metody na wielokrotne powtórzenia słów we fragmentach tekstu. Fragmenty dla których prawdopodobieństwo przynależności do zdefiniowanych kategorii jest mniejsze od ustalonego progu są następnie odrzucane, zaś pozostałe grupowane są algorytmem aglomeracyjnym. W ten sposób możliwe jest szybkie zidentyfikowanie najbardziej reprezentatywnych elementów opisu produktów w zbiorze stron WWW, które odpowiadają zadanym słowom kluczowym; opisy te mogą być następnie przedmiotem dalszego przetwarzania w

[6] Zadanie badawcze SYNAT (System Nauki i Techniki) pt. „*Utworzenie uniwersalnej, otwartej, repozytoryjnej platformy hostingowej i komunikacyjnej dla sieciowych zasobów wiedzy dla nauki, edukacji i otwartego społeczeństwa wiedzy*", grant nr SP/I/1/77065/10 Narodowego Centrum Badań i Rozwoju

systemie automatycznym lub też mogą być bezpośrednio wykorzystane przy tworzeniu baz danych serwisów e-commerce.

Z kolei problem analizy opinii użytkowników (ang. *opinion-mining*) jest przedmiotem prowadzonych obecnie przez habilitanta badań, których opis zawiera artykuł [P11]. Badania te doprowadziły do nawiązania współpracy pomiędzy Instytutem Informatyki Politechniki Warszawskiej, a Grupą Allegro. Ich kontynuacją jest prowadzony obecnie przez habilitanta projekt, mający na celu analizę danych zbieranych w największych polskich serwisach *e-commerce*, w tym w szczególności *allegro.pl* i *ceneo.pl.* Celem projektu jest m.in. opracowanie nowych algorytmów rekomendacji produktów, które są następnie wdrażane w wyżej wspomnianych serwisach[7].

d) Przetwarzanie i wizualizacja informacji heterogenicznej

Heterogeniczna natura informacji dostępnej w Internecie sprawia, iż zastosowanie do jego analizy i pozyskiwania zeń ustrukturalizowanej informacji semantycznej wyłącznie narzędzi opartych o metody przetwarzania języka naturalnego jest nieefektywne. Internet jest medium, znajdującym się na pograniczu systemów o w pełni kontrolowanej strukturze, takich jak bazy relacyjne i systemów nieustrukturalizowanych, takich jak korpusy dokumentów tekstowych. Nie będąc klasyczną bazą dokumentów tekstowych, udostępnia jednak pewne metadane, dotyczące tematyki zasobów czy ich popularności, choć ich wydobycie wymaga często dodatkowego przetwarzania. Ponadto, pewne podzbiory globalnej sieci charakteryzują wewnętrzną, uporządkowaną strukturą.

Przykładem takiego podzbioru, są serwisy typu *wikiwiki*, w tym w szczególności Wikipedia. Ich cechą charakterystyczną jest łatwość modyfikacji zawartości przez użytkowników bezpośrednio za pomocą przeglądarki internetowej, zwykle za pomocą specjalizowanego języka znaczników, bądź też dedykowanego edytora działającego w środowisku przeglądarki. Pozwala to na szybkie tworzenie i modyfikację treści tekstowych w modelu społecznościowym, jednocześnie przez wielu użytkowników sieci. Dotyczy to jednak wyłącznie danych tekstowych. W połowie ubiegłej dekady, gdy tego typu serwisy społecznościowe gwałtownie zyskiwały na popularności, obróbka innego rodzaju danych, w tym w szczególności graficznych, była właściwie niemożliwa. Oczywiście serwisy tego typu ewoluowały, jednak powstające modyfikacje dotyczyły, jak to przedstawia Cunningham [22] jedynie ulepszeń mechanizmów edycji tekstu, bądź też tworzone systemy – np. [23], [24] okazywały się trudne w użyciu, wymagając bądź to instalacji dodatkowego oprogramowania, bądź też specjalistycznej wiedzy. Tym samym społecznościowe tworzenie publicznie dostępnych zasobów wiedzy było utrudnione, wiele bowiem wartościowych informacji daje się wyrazić w zwięzłej formie jedynie w postaci nietekstowej, dotyczy to w szczególności różnego rodzaju diagramów (np. urządzeń technicznych czy procesów) oraz map.

Wspomniane powyżej ograniczenie było inspiracją dla opracowania przez habilitanta metodyki tworzenia map w modelu społecznościowym, możliwym do wykorzystania w istniejących systemach typu *wiki*. Pozwala ona na szybkie tworzenie zasobów informacji geograficznej o jakości wystarczającej dla większości zastosowań praktycznych (takich jak nawigacja, tworzenie map itp.), pomimo, iż poszczególni społeczności mogą nie dysponować dokładną informacją geodezyjną. Dzięki temu możliwe jest uniknięcie wstępnej, powolnej fazy wzrostu systemu, wynikającej z jego

[7] Patrz załączone referencje od przedstawiciela firmy Grupa Allegro sp. z o.o.

początkowej niewielkiej atrakcyjności dla użytkowników nie będących specjalistami informacji geograficznej. Przykładem jest tu początkowa ewolucja serwisu OpenStreetMap [23] przedstawiona w artykule [P3], zawierającym także opis metodyki zaproponowanej przez habilitanta. Prototyp systemu ją realizującego, zbudowany przy wykorzystaniu dopiero zyskującej wówczas popularność technologii AJAX [25] i bezpośredniego obrazowania obiektów wektorowych przez przeglądarkę internetową, przedstawiony został przez habilitanta na światowej konferencji fundacji Wikimedia na Harvard University, w 2006 roku[8]. Swego rodzaju potwierdzeniem trafności tego rodzaju podejścia do akwizycji danych geograficznych mogą być, powstałe później, serwisy takie jak Waze[9], których zasada działania jest podobna, choć oczywiście szczegóły implementacji są odmienne.

Tego rodzaju systemy informatyczne, pozwalające na tworzenie treści przez społeczności użytkowników w modelu otwartym, w tym w szczególności Wikipedia, stawały się wówczas jednymi z najważniejszych narzędzi debaty publicznej. Jednocześnie ta sieciowa encyklopedia stała się jednym z najistotniejszych źródeł informacji semantycznej dostępnym w sieci WWW. Jest to zasób, który określić można mianem semi-ustrukturalizowanego. Większość danych w nim zawartych dostępna jest jedynie, podobnie jak w innych systemach będących przedmiotem dotychczasowych prac badawczych habilitanta, w postaci zapisu w języku naturalnym. Jednocześnie jednak istnieje bogata warstwa metadanych. Są one zapisane *explicite* jak to ma miejsce w przypadku niektórych typów obiektów takich jak osoby czy lokalizacje geograficzne. Mogą być też dostępne nie wprost, dzięki analizie kategorii artykułów, powiązań między nimi i wreszcie historii edycji danych, której analiza wydaje się być szczególnie istotna ze względu na społecznościowy charakter Wikipedii. Rozmiar i dynamika zmian treści Wikipedii powodują jednak, iż do obserwacji zachodzących wśród jej użytkowników procesów niezbędne są odpowiednie narzędzia - analityczne i wizualizacji danych, przy czym klasyczne metody wizualizacji zawartości statycznych dokumentów wieloautorskich zaproponowane np. w [26], [27] czy [28] nie pozwalają na odzwierciedlenie zmian zachodzących w systemie typu *wiki*. Jedyne istniejące w tym czasie narzędzie, służące wizualizacji dynamiki Wikipedii – History Flow F. Viegas'a [29] dedykowane jest analizie języka angielskiego, ponadto zaś umożliwia wygodną obserwację jedynie wielkoskalowych zmian. Narzędzie, pozwalające na analizę historii także niewielkich edycji zostało zaimplementowane przez habilitanta w 2007. Było ono następnie wykorzystane do wyszukania i zobrazowania interesujących trendów w historii edycji artykułów polskiej wersji Wikipedii, takich jak konflikty pomiędzy autorami czy też zjawisko „wandalizmu edycyjnego”, co opisano w artykule [P4]. Wydaje się, iż była to pierwsza próba wizualizacji polskiej wersji językowej Wikipedii. Jednym z ubocznych wyników prowadzonych wówczas badań było spostrzeżenie, iż istnieje bezpośredni związek pomiędzy częstością uaktualnień danego artykułu w Wikipedii, a popularnością związanego z nim pojęcia w świadomości społecznej. Informacja ta została wykorzystana w kolejnym projekcie badawczym habilitanta, wykonanym podczas stażu na Uniwersytecie Stanforda, realizowanym w ramach programu Top 500 Innovators MNiSW, w ramach którego opracowano metodę i narzędzie generalizacji map geograficznych. Problem generalizacji map, tj. odpowiedniego jej uproszczenia w zależności od skali oraz jakości typograficznej mapy, tak aby zwiększyć jej czytelność, jest zadaniem wymagającym

[8] Gawrysiak Piotr: Plans not maps! Wikiplan – collaborative geographical map building system, In: 2nd Wikipedia Conference Wikimania 2006, Cambridge, Boston, USA, 2006

[9] Serwis Waze [41], powstały w Izraelu w 2008 roku, został w 2013 roku kupiony przez firmę Google za sumę ok. 1 miliarda dolarów.

specjalistycznej wiedzy i wciąż wykonywanym ręcznie. Automatyzacja tego procesu jest przedmiotem aktualnie prowadzonych badań – przy czym o ile osiągnięto pewne obiecujące rezultaty odnośnie upraszczania kształtów obiektów ciągłych (takich jak linie brzegowe czy granice państw – patrz [30], [31]), to wybór najbardziej istotnych obiektów na mapie, wciąż stanowi istotny problem i dokonywany jest najczęściej w oparciu o pewne statyczne cechy ww. obiektów (jak np. liczba mieszkańców miasta) [32]. Tymczasem w metodzie zaproponowanej przez habilitanta wykorzystano między innymi aktywność użytkowników Wikipedii do uwypuklenia najbardziej istotnych w danym czasie obiektów na mapie. W tym celu zdefiniowane zostały dwie miary istotności obiektów. Miara pasywnej popularności (*ang. passive popularity*) jest uzależniona od liczby zasobów sieciowych dotyczących obiektu i aproksymowana jest poprzez analizę frekwencyjną unigramowej reprezentacji korpusu stron WWW oraz zasobów serwisów społecznościowych, takich jak Twitter. Z kolei wartość miary aktywnej popularności (ang. *active popularity*) jest proporcjonalna do liczby dostępów (*ang. access*) do danego zasobu w analizowanej jednostce czasu, przy czym przez dostęp rozumiane jest tu zarówno pobranie zawartości zasobu przez użytkownika, jak też i jego cytowanie, np. w postaci opublikowania hiperłącza. Wartości obu miar są następnie wykorzystywane przy tworzeniu wynikowej mapy, bądź to poprzez wykorzystanie reprezentacji barwnej, bądź też poprzez pomijanie tych obiektów, dla których wartości miar znajdują się poniżej wartości progowej (patrz przykład w artykule [P9]).

Problem wizualizacji informacji pojawia się jeszcze kilkakrotnie w pracach badawczych prowadzonych przez habilitanta. Jednym z eksperymentów wykonanych we wspomnianym w poprzednim punkcie projekcie SYNAT była analiza baz danych prowadzonych przez Ośrodek Przetwarzania Informacji (m.in. baza „Nauka Polska” [33]), której celem było zbudowanie grafu powiązań pomiędzy polskimi naukowcami ze stopniem naukowym doktora habilitowanego, wynikających ze współpracy badawczej, następnie zaś wykorzystanie tej informacji do zidentyfikowania stopnia kooperacji pomiędzy polskimi ośrodkami naukowymi. Jednym z wyników była wizualizacja geograficzna (inspirowana wizualizacją geograficzną sieci Facebook [34]) sieci współpracy polskich uniwersytetów i innych ośrodków badawczych, zaprezentowana m.in. w artykule [P7]. Nadmienić przy tym należy, iż akwizycja informacji na potrzeby tego eksperymentu wymagała wykorzystania opracowanych w tym celu metod *web mining*, jako że Ośrodek Przetwarzania Informacji nie zapewnił bezpośredniego dostępu do wspomnianych baz.

Uzupełniającym spojrzeniem na tematykę przedstawionych powyżej prac badawczych habilitanta, może być ich charakterystyka za pomocą następujących słów kluczowych:

- ontologie / sieci semantyczne,
- przetwarzanie tekstu,
- uczenie maszynowe,
- wizualizacja informacji.

W pracach habilitanta wyróżnić przy tym można dwa splatające się i uzupełniające się wątki: *par excellence* naukowy, obejmujący badania algorytmów i metod przetwarzania informacji, oraz

technologiczno-wdrożeniowy, związany z budową narzędzi informatycznych nadających się do praktycznego zastosowania. Poniższa tabela podsumowuje tematykę tych prac.

Nr publikacji	Ontologie / sieci semantyczne	Przetwarzanie tekstu	Uczenie maszynowe	Wizualizacja informacji	Inżynieria oprogramowania
1		+	+		
2		+	+		+
3				+	+
4		+		+	
5	+	+	+		+
6	+	+	+		+
7		+		+	
8	+	+	+		
9		+	+	+	
10	+				+
11	+	+	+		

5. Omówienie pozostałych osiągnięć naukowo-badawczych

5.1 Charakterystyka osiągnięć naukowo-badawczych uzyskanych w toku realizacji prac badawczych

Rozwój informatyki w ostatnim dziesięcioleciu, w tym w szczególności rozwój sieci Internet, doprowadził do sytuacji w której właściwie wszystkie zasoby ludzkiej wiedzy dostępne są już w postaci cyfrowej. Potencjalnie osiągalne są one tym samym z niemalże z dowolnego miejsca na Ziemi. Ich cyfryzacja i upublicznienie w sieci nie oznaczają jednak, iż są łatwe do odnalezienia, czy też wykorzystania – w tym w szczególności przez autonomiczne systemy informatyczne, które wciąż nie są w stanie, pomimo rozwoju badań nad sztuczną inteligencją, posługiwać się bezpośrednio wiedzą zapisaną w języku naturalnym. Zasoby sieciowe wymagają opisania poprzez metadane, interpretacji i konwersji do bardziej formalnych systemów reprezentacji wiedzy. Tworzenie zaś, możliwe automatycznych, metod to umożliwiających jest przedmiotem opisanych powyżej badań habilitanta. Główną myślą przewodnią tych badań było kreowanie zasobów wiedzy w postaci ustrukturalizowanej, nadającej się do dalszego praktycznego wykorzystania – czy to w sposób formalny, czy też w wyniku zastosowania odpowiedniej wizualizacji – z zasobów nieuporządkowanych, łatwych do odnalezienia, ale niezbyt przydatnych dla systemów informatycznych. Przywołując znaną w środowisku eksploracji danych metaforę piramidy informacyjnej, w badaniach tych powstawały narzędzia ułatwiające poruszanie się w górę owej piramidy, w stronę wiedzy. Oczywiście niezależnie od powodzenia wysiłków nad stworzeniem systemów sztucznej inteligencji, aby nawiązać tu do myśli sformułowanej we wstępie do niniejszego autoreferatu, narzędzia te są już obecnie bezpośrednio użyteczne dla ludzi zajmujących się wyszukiwaniem i obróbką informacji. Większość z tych narzędzi i metod została opisana

powyżej, jako część osiągnięcia naukowego habilitanta.

Niektóre z rozwiązań opracowanych przez habilitanta wykorzystane zostały zaś w innych projektach badawczych, nie należących do osiągnięcia naukowego. Wśród nich należy wymienić w szczególności projekt 6-go programu ramowego Komisji Europejskiej KNOW-IT, w ramach którego habilitant przebywał w 2007 roku w Rzymie w firmie CiaoTech S.r.l jako *visiting researcher*. W projekcie tym tworzono metodykę organizacji prac B+R w małych przedsiębiorstwach opartą na założeniach metody TRIZ [35]. Częścią prowadzonych prac badawczych była próba zastosowania algorytmów eksploracji danych do odkrywania trendów (takich jak np. powstawanie i rozpowszechnienie wykorzystania nowych technologii) w bazach patentowych Europejskiego Urzędu Patentowego. Skrótowy opis wspomnianego projektu zawiera artykuł [36].

Problem reprezentacji, tworzenia i wykorzystywania zasobów wiedzy ma oczywiście także niebagatelne znaczenie dla nauk humanistycznych. Nie jest to zresztą odosobniony przypadek w informatyce. Przykładem mogą być prace profesora Zdzisława Pawlaka, który opracowawszy teorię zbiorów przybliżonych (ang. *rough sets*) jako narzędzie eksploracji danych, znajdował następnie jej zastosowania np. w naukach prawniczych [37] oraz filozofii [38]. Aspekty, w szczególności jakościowe, przetwarzania zasobów wiedzy dotyczące nauk humanistycznych habilitant analizował m.in. w swej rozprawie habilitacyjnej obronionej w 2010 roku na Wydziale Historycznym Uniwersytetu Warszawskiego [39]. W dziedzinie nauk technicznych badania habilitanta prowadzą jednak znacznie dalej, zaś ich celem nie jest jedynie wskazanie obszarów zastosowań metod automatycznego przetwarzania wiedzy, ale przede wszystkim zaproponowanie konkretnych algorytmów i heurystyk, oraz narzędzi, które pozwalają na strukturalizację zasobów wiedzy, tak by nadawała się do przetwarzania maszynowego.

Do innych obszarów badań habilitanta w dziedzinie nauk technicznych, nie należących do przedstawionego osiągnięcia naukowego, zaliczyć można w szczególności rozwój szeroko rozumianych systemów mobilnych. Habilitant prowadził szereg projektów przemysłowych, realizowanych w Instytucie Informatyki PW lub też w Laboratorium Technologii Mobilnych BRAMA Polskiej Telefonii Cyfrowej i Politechniki Warszawskiej, związanych z tworzeniem oprogramowania działającego na urządzeniach mobilnych. Prace te w szczególności dotyczyły innowacyjnych rozwiązań mobilnych interfejsów użytkownika. Wymienić tu można projekt mobilnego interfejsu dla urządzeń mobilnych oparty o technologię Google Gears, realizowany dla firmy Samsung; czy też prace nad stworzeniem systemów nawigacji wykorzystujących paradygmat rzeczywistości rozszerzonej. Ten drugi temat realizowany był m.in. w ramach międzynarodowego projektu badawczego CRUMBS Celtic Plus przez kierowany przez habilitanta zespół współpracującej z Politechniką Warszawską firmy Polidea. Habilitant prowadzi także nieprzerwanie od 2008 roku wykład – w początkowym okresie unikalny w skali krajowej – poświęcony projektowaniu i tworzeniu aplikacji dla urządzeń mobilnych.

W pierwszych latach po uzyskaniu stopnia naukowego doktora zainteresowania naukowe habilitanta obejmowały także problematykę klasycznej (tj. nie dotyczącej analizy języka naturalnego) eksploracji danych. Związane było to między innymi ze wspomnianymi powyżej projektami badawczymi prowadzonymi dla polskich operatorów telefonii komórkowej. W obszarze zainteresowań habilitanta znalazły się także, w wyniku współpracy nawiązanej z Uniwersytetem w

Antwerpii, badania dotyczące zastosowania metod *text mining* do analizy informacji genetycznej. W ostatnim okresie habilitant powrócił do tej tematyki organizując, wraz z Michałem Okoniewskim z ETH w Zurichu, zespół badawczy bioinformatyki w ramach Zakładu Systemów Informacyjnych Instytutu Informatyki, którego głównym obszarem zainteresowań jest obecnie analiza danych pochodzących z systemów sekwencjonowania DNA i RNA nowej generacji (patrz np. artykuł [40]).

5.2 Osiagnięcia naukowo-badawcze – informacje bibliometryczne

Dorobek publikacyjny habilitanta obejmuje 50 publikacji (46 napisanych po uzyskaniu stopnia doktora) w tym 2 monografie, 3 książki redagowane i 45 referatów w recenzowanych materiałach konferencyjnych i czasopismach, sumaryczna liczba punktów MNiSW ww. publikacji wynosi 301. Według bazy danych Google Scholar sumaryczna liczba cytowań publikacji habilitanta to 96, zaś indeks Hirsh'a obliczony według tej samej bazy wynosi 6.

6. Literatura dodatkowa

[1] Penrose R.: The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics, 1989, Oxford University Press, Inc., New York, NY, USA

[2] McCarthy J.: Defending AI research : a collection of essays and reviews., in: CSLI lecture notes: no. 49. Center for the Study of Language and Information, 1996. Cambridge University Press, UK

[3] McCallum A., Nigam K., A Comparision of Event Models for Naive Bayes Text Classification, AAAI-98 Workshop on Learning for Text Categorization, 1998

[4] Yang Y., Liu X., A re-examination of text categorization methods, ACM SIGIR Conference on Research and Development in Information Retrieval, New York, 1998

[5] M.F. Porter, An algorithm for suffix stripping, *Program*, 14 (3) pp 130−137, 1980

[6] Dębowski Ł., Trigram morphosyntactic tagger for Polish. W: Intelligent Information Processing and Web Mining. IIS:IIPWM'04, Zakopane, Poland, 409-413. Springer Verlag, 2004

[7] Weiss D., A Survey of Freely Available Polish Stemmers and Evaluation of Their Applicability in Information Retrieval. W: Zygmunt Vetulani (red.), Proceedings of the 2nd Language & Technology Conference, s. 216-221, Poznań, 2005

[8] Berners-Lee T. et al.: The Semantic Web, in: Scientific American 284 , no. 5, 2001, pp. 34-43

[9] Brin S., Page L.: The anatomy of a large-scale hypertextual Web search engine., in: Proceedings of the seventh international conference on World Wide Web 7 (WWW7), 1998, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, pp. 107-117

[10] Marchionini G.: Exploratory search: from finding to understanding., in: Commun. of the ACM 49, 4 (April 2006), pp. 41-46

[11] Hackathorn R.D.: Web Farming for the Data Warehouse, 1999, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA

[12] John H. Gennari et al., The evolution of Protégé: an environment for knowledge-based systems development, International Journal of Human-Computer Studies, Volume 58, Issue 1 , pp. 89-123,

2003

[13] Maedche, A., Staab, S.: Mining Ontologies from Text. In: Dieng, R., Corby, O.(eds.) EKAW 2000. LNCS (LNAI), vol. 1937, pp. 189–202. Springer, Heidelberg, 2000

[14] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In Journal of the ACM, 46(5), pages 604-632, 1999

[15] Hamon, T., Nazarenko, A., Gros, C.: A step towards the detection of semantic variants of terms in technical documents. In: Proc. 36 Ann. Meeting of ACL, 1998

[16] Wu, H., Zhou, M.: Optimizing Synonym Extraction Using Monolingual and Bilingual Resources. In: Ann. Meeting ACL, Proc. 2nd Int'l Workshop on Paraphrasing, vol. 16, pp. 72–79, 2003

[17] Hammer, J., Garcia-molina, H., Cho, J., Aranha, R., Crespo, A.: Extracting semistructured information from the web. In: In Proceedings of the Workshop on Management of Semistructured Data, pp. 18–25, 1997

[18] Liu, L., Pu, C., Han, W.: XWRAP: An XML-enabled wrapper construction system for web information sources. In: Proceedings of the 16th International Conference on Data Engineering, pp. 611–621. IEEE Computer Society, Washington, DC, USA, 2000

[19] Chang, C.H., Lui, S.C.: IEPAD: information extraction based on pattern discovery. In: Proceedings of the 10th International Conference on World Wide Web, WWW 2001, pp. 681–688. ACM, New York, 2001

[20] Crescenzi, V., Mecca, G., Merialdo, P.: RoadRunner: Towards automatic data extraction from large web sites. In: Proceedings of the 27th International Conference on Very Large Data Bases, VLDB 2001, pp. 109–118. Morgan Kaufmann Publishers Inc., San Francisco, 2001

[21] Zhai, Y., Liu, B.: Extracting web data using instance-based learning. World Wide Web 10, 113–132, 2007

[22] W. Cunningham, Wikis Then and Now, Intl. Wikimania Conference, Frankfurt, 2005

[23] http://www.openstreetmap.org, dostęp 17 listopada 2014

[24] Haraguchi Y., et al, The Living Map-A communication tool that connects real world and online community by using a map, Journal of the Asian Design International Conference, 2003

[25] Garrett J.J., „Ajax: A New Approach to Web Applications", Adaptive Path, 2005, https://web.archive.org/web/20080702075113/http://www.adaptivepath.com/ideas/essays/archives/000385.php, dostęp 17 listopada 2014

[26] Pennock K et al. „Visualizing the non-visual: spatial analysis and interaction with information from text documents". In Proc. on Information Visualization, 1995.

[27] M. Brewster H. Foote N. E. Miller, P. C. Wong. Topic islands - a wavelet-based text visualization system. In IEEE Visualization, Proc. of the Conference on Visualization, 1998

[28] H. Lieberman H. Liu, T. Selker. Visualizing the affective structure of a text document. In: Conference on Human Factors in Computing Systems, 2003

[29] K. Dave F. Viegas, M. Wattenberg. Studying cooperation and conflict between authors with

history flow visualizations. In Proc. of SIGCHI, 2004

[30] Grunreich, D., Development of computer-assisted generalization on the basis of cartographic model theory, In: GIS and Generalization: Methodology and Practice. Müller, J.P. Lagrange, R. Weibel. Bristol (eds.), Taylor & Francis, pp. 47-55, 1995

[31] Foerster, T., J. E. Stoter, and M. Kraak, Challenges for Automated Generalisation at European Mapping Agencies: a Qualitative and Quantitative Analysis, In: The Cartographic Journal 47, No. 1, pp. 41–54, 2010

[32] Duchêne, C. Automated Map Generalisation Using Communicating Agents, In: Proceedings of the 21st International Cartographic Conference (ICC), Durban, South Africa, 2003

[33] http://www.nauka-polska.pl/Ludzie-nauki.html, dostęp 17 listopada 2014

[34] Butler P.: Visualizing Friendships, Facebook, USA, 2010 https://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

[35] Altshuller G.S.: Creativity as an exact science, 1984, Gordon and Breach Science Publ., Amsterdam, The Nederlands

[36] Gawrysiak P.: KNOW-IT: Methodology and software for driving innovation processes in small and medium enterprises, in: 4th European Conference on Management Leadership and Governance, 2008, Reading, UK, pp. 47-49

[37] Pawlak, Z.: A primer on rough sets: A new approach to drawing conclusions from data, in: Cardozo Law Review, 22(5-6):1407-1415, 2002

[38] Pawlak, Z.: Orthodox and Non-orthodox Sets - some Philosophical Remarks. Foundations of Computing and Decision Sciences, 30(2): 133-140, 2005

[39] Gawrysiak P.: Cyfrowa rewolucja. Rozwój cywilizacji informacyjnej, 2008, Wydawnictwo Naukowe PWN SA, ISBN 978-83-01-15607-7

[40] Wiewiórka M., Messina A., Pacholewska A., Maffioletti S., Gawrysiak P., Okoniewski M.: SparkSeq: fast, scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision, in: Bioinformatics. 2014 Sep 15; 30(18):2652-3. doi: 10.1093/bioinformatics/btu34

[41] http://www.waze.com, dostęp 17 listopada 2014