Autoreferat

przedstawiający opis podstawowego osiągnięcia dorobku naukowego
w związku z ubieganiem się o nadanie stopnia doktora habilitowanego

dr Marek Z. Reformat

Wydział Inżynierii Elektrycznej i Komputerowej
Uniwersytet Alberta
Edmonton, Alberta, T6G 1H9, Kanada
tel: 780.492.2848
e-mail: Marek.Reformat@ualberta.ca
`http://www.ualberta.ca/~reformat`

I Imię i Nazwisko

- Marek Z. Reformat

II Posiadane dyplomy i stopnie naukowe

- **Doktorat: Wydział Inżynierii Elektrycznej i Komputerowej, 1997.**
  Uniwersytet Manitoba.
  Rozprawa: Aplikacja Zaawansowanego Kompensatora Mocy Biernej
  (tytuł oryginalny: Application of Advanced Static VAR Compensator at ac/dc interconnection).
  Promotor: Dr. E. Kuffel.

- **Mgr inż. (z wyróżnieniem): Wydzial Elektryczny, 1988.**
  Politechnika Poznańska.
  Praca dyplomowa: Mikroprocesorowa Implementacja Protokołów Prezentacji i Wirtualnego Terminala.
  Promotor: Dr. W. Cellary

III Zatrudnienie

- **Profesor, lipiec 2008 – obecnie.**
  Wydział Inżynierii Elektrycznej i Komputerowej, Uniwersytet Alberta, Kanada
  (Wicedyrektor Instytutu do Spraw Studentów: Magistrów i Doktorantów: 09.2015 – obecnie)
  (Dyrektor programu Inżynierii Komputerowej: 09.2010 – 09.2015)
- **Associate Profesor, lipiec 2004 – czerwiec 2008.**
  Wydział Inżynierii Elektrycznej i Komputerowej, Uniwersytet Alberta, Kanada
- **Assistant Profesor, lipiec 2000 – czerwiec 2004.**
  Wydział Inżynierii Elektrycznej i Komputerowej, Uniwersytet Alberta, Kanada

- **Distribution System Researcher, grudzień 1996 – czerwiec 2000.**
  Manitoba HVDC Research Centre, Winnipeg, Kanada
- **Projektant Sieci Komputerowych, wrzesień 1988 – grudzień 1992.**
  AdvaCom Ltd., Poznań, Polska
- **Pracownik badawczy, wrzesień 1988 – grudzień 1992.**
  Instytut Informatyki, Politechnika Poznańska, Polska
- **Praktyka studencka, czerwiec 1987.**
  EI-Honeywell, Nis, Jugoslawia.

IV Wskazane osiągnięcia naukowe

Jako osiągnięcia naukowe, wynikające 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. 2016 r. poz. 882, ze zm. w Dz. U. z 2016 r. poz. 1311) podaję jednotematyczny cykl publikacji zatytułowany:

Metody Inteligencji Obliczeniowej
w Inteligentnych Systemach Webowych

IV.1 Cykl publikacji wchodzących w skład osiągnięcia

Cykl zawiera dwanaście tytułów, które zostały wybrane z moich publikacji. Lista publikacji cyklu podana jest poniżej.

(R1) P.D. Hossein Zadeh[1](40%), M. Hossein Zadeh (20%), **M. Reformat** (40%), Feature-driven Linguistic-based Entity Matching in Linked Data with Application in Pharmacy, *Soft Computing*, Vol. 21, No. 2, 2017, str. 353-368. **IF 2.365[2]; 25 pkt.**[3]
Mój całkowity udział w przygotowaniu publikacji szacuję na 40%, w szczególności: koncepcja i sformułowanie problemu 75%; projekt i implementacja rozwiązania 75%; eksperymenty, analiza i interpretacja wyników 10%; przygotowanie publikacji 65%.

(R2) P.D. Hossein Zadeh (50%), **M. Reformat** (50%), Assessment of Semantic Similarity of Concepts Defined in Ontology, *Information Sciences*, Vol. 250, No, 20, 2013, str. 21-39. **IF 4.305; 45 pkt.**
Mój całkowity udział w przygotowaniu publikacji szacuję na 50%, w szczególności: koncepcja i sformułowanie problemu 65%; projekt i implementacja rozwiązania 65%; eksperymenty, analiza i interpretacja wyników 20%; przygotowanie publikacji 65%.

(R3) R. R. Yager (40%), **M. Reformat** (60%), Looking for Like-minded Individuals in Social Networks using Tagging and Fuzzy Sets, *IEEE Transactions on Fuzzy Systems*, Vol. 21, No. 4, 2013, str. 672-687. **IF 8.415; 50 pkt.**
Mój całkowity udział w przygotowaniu publikacji szacuję na 60%, w szczególności: koncepcja i sformułowanie problemu 50%; projekt i implementacja rozwiązania 50%; eksperymenty, analiza i interpretacja wyników 90%; przygotowanie publikacji 85%.

[1] Podkreślone nazwisko oznacza studenta.
[2] Impact Factor z roku 2017/2018.
[3] Punktacja według wykazu M.N.i Sz. W. z 26-01-2017 (https://punktacjaczasopism.pl).

(R4) P.D. Hossein Zadeh (30%), **M. Reformat** (70%), Context-aware Similarity Assessment within semantic Space formed in Linked Data, *Journal of Ambient Intelligence and Humanized Computing*, Vol. 4, No. 4, 2013, str. 515-532. **IF 1.423; 20 pkt.**
Mój całkowity udział w przygotowaniu publikacji szacuję na 70%, w szczególności: koncepcja i sformułowanie problemu 80%; projekt i implementacja rozwiązania 80%; eksperymenty, analiza i interpretacja wyników 50%; przygotowanie publikacji 80%.

(R5) **M. Reformat** (50%), R.R. Yager (15%), Z. Li (30%), N. Alajlan (5%), Human-inspired Identification of High-level Concepts using OWA and Linguistic Quantifiers, *International Journal of Computers, Communications and Control*, Vol. VI, No. 3, 2011, str. 473-502. **IF 1.290; 20 pkt.**
Mój całkowity udział w przygotowaniu publikacji szacuję na 50%, w szczególności: koncepcja i sformułowanie problemu 60%; projekt i implementacja rozwiązania 40%; eksperymenty, analiza i interpretacja wyników 20%; przygotowanie publikacji 50%.

(R6) R.R. Yager (25%), **M. Reformat** (50%), G. Gumrah (25%), Using a Web Personal Evaluation Tool - PET for Lexicographic Multi-criteria Service Selection, *Knowledge-Based Systems*, Vol. 24, No. 7, 2011, str. 929-942. **IF 4.396; 40 pkt.**
Mój całkowity udział w przygotowaniu publikacji szacuję na 50%, w szczególności: koncepcja i sformułowanie problemu 45% ; projekt i implementacja rozwiązania 35%; eksperymenty, analiza i interpretacja wyników 10%; przygotowanie publikacji 70%.

(R7) **M. Reformat** (60%), and R.R. Yager (40%), Tag-based Fuzzy Sets for Criteria Evaluation in On-line Selection Processes, *Journal of Ambient Intelligence and Humanized Computing*, Vol. 2, No. 1, 2011, str. 35-51. **IF 1.423; 20 pkt.**
Mój całkowity udział w przygotowaniu publikacji szacuję na 60%, w szczególności: koncepcja i sformułowanie problemu 60%; projekt i implementacja rozwiązania 60%; eksperymenty, analiza i interpretacja wyników 90%; przygotowanie publikacji 90%.

(R8) **M. Reformat** (60%), and S. K. Golmohammadi (40%), Rule- and OWA-based Semantic Similarity for User Profiling, *International Journal of Fuzzy Systems*, Vol. 12, No. 2, 2010, str. 87-102. **IF 2.396; 25 pkt.**
Mój całkowity udział w przygotowaniu publikacji szacuję na 60%, w szczególności: koncepcja i sformułowanie problemu 70%; projekt i implementacja rozwiązania 60%; eksperymenty, analiza i interpretacja wyników 25%; przygotowanie publikacji 60%.

(R9) Z. Li (50%), and **M. Reformat** (50%), A Schema for Ontology-based Concept Definition and Identification, *International Journal of Computer Applications in Technology*, Vol. 38, No. 4, 2010, str. 333-345. **IF 0.758; -.**
Mój całkowity udział w przygotowaniu publikacji szacuję na 50%, w szczególności: koncepcja i sformułowanie problemu 75%; projekt i implementacja rozwiązania 25%; eksperymenty, analiza i interpretacja wyników 20%; przygotowanie publikacji 80%.

(R10) R.R. Yager (30%), **M. Reformat** (50%), G. Gumrah (20%), WebPET: An Online Tool for Lexicographic Decision Making, *IEEE Intelligent Systems*, Vol. 25, No. 6, 2010, str. 76-83. **IF 2.596; 40 pkt.**
Mój całkowity udział w przygotowaniu publikacji szacuję na 50%, w szczególności: koncepcja i sformułowanie problemu 45%; projekt i implementacja rozwiązania 35%; eksperymenty, analiza i interpretacja wyników 15%; przygotowanie publikacji 85%.

(R11) **M. Reformat** (50%), R.R. Yager (20%), and Z. Li (30%), Ontology Enhanced Concept Hierarchies for Text Identification, *International Journal on Semantic Web and Information Systems*, Vol. 4, No.

3, 2008, str. 16-43. **IF 0.793; 40 pkt.**
Mój całkowity udział w przygotowaniu publikacji szacuję na 50%, w szczególności: koncepcja i sformułowanie problemu 35%; projekt i implementacja rozwiązania 70%; eksperymenty, analiza i interpretacja wyników 25%; przygotowanie publikacji 65%.

(R12) **M. Reformat** (60%), and C. Ly (40%), Ontological Approach to Development of Computing with Words based Systems, *International Journal of Approximate Reasoning*, Vol. 50, No. 1, 2009, str. 72-91. **IF 1.767; 35 pkt.**
Mój całkowity udział w przygotowaniu publikacji szacuję na 60%, w szczególności: koncepcja i sformułowanie problemu 70%; projekt i implementacja rozwiązania 60%; eksperymenty, analiza i interpretacja wyników 20%; przygotowanie publikacji 70%.

IV.2 Cel naukowy cyklu publikacji i znaczenie osiągniętych wyników

Obecny stan sieci internetowej wydaje się bardzo atrakcyjny – zakres gromadzonej informacji wzrasta, użytkownicy mają nieograniczony dostęp do danych, dostarczone im informacje dobrane są do ich zainteresowań i potrzeb. Bliższe spojrzenie na proces selekcji danych wywołuje jednak pewien niepokój. Web wydaje się lustrzanym odbiciem aktywności użytkowników: dostarczane im informacje zależą od poszukiwanych wcześniej na sieci danych, odwiedzanych portali i od tego, co było dla nich interesujące i ważne. Sytuacja taka oznacza, że użytkownicy mają coraz mniejszą możliwość dotarcia do nowych, potencjalnie interesujących i atrakcyjnych informacji.

W tym kontekście, prowadzone przeze mnie badania mają na celu podniesienie jakości korzystania z weba, to znaczy zapewnienie użytkownikom dostępu do informacji, która jest dla nich wyselekcjonowana i wstępnie przetworzona, wzbogacającej ich wiedzę. Celem badań jest wyjście poza tradycyjny mechanizm wyszukiwania danych oparty na słowach kluczowych. Prowadzi to również do konstrukcji struktur danych wymaganych do inteligentnego i przyjaznego dla użytkownika (ang. human-friendly) korzystania z weba.

Przedstawione w autoreferacie badania dotyczą zastosowań technik i metodologii inteligencji obliczeniowej do pełniejszego wykorzystania danych dostępnych w sieciach komputerowych. Koncentrują się one na czterech tematach: analizie danych z **portali społecznościowych** (ang. social networks), konstrukcji **profili użytkowników i systemów rekomenduyących** (ang. recommender systems), budowie **algorytmów do wzbogacania informacji w oparciu o ontologie** oraz **przetwarzaniu danych przedstawionych w formacie RDF** (ang. Resource Description Framework).

Zakres badań dotyczących **portali społecznościowych** obejmuje automatyczną konstrukcję sygnatur pojedyńczych użytkowników oraz grup użytkowników. Owe sygnatury odzwierciedlają zainteresowania i aktywności użytkowników i są budowane w oparciu o mechanizmy i operacje znane ze zbiorów rozmytych. Istnienie sygnatur pozwala na wyszukiwanie użytkowników i grup, które spełniają wymagania podobieństwa do stopnia ustalonego przez użytkowaników.

Budowa **profili użytkowników i systemów rekomendacyjnych** to następny temat moich badań naukowych. W tym przypadku, profile tworzone są w oparciu o analizę aktywności użytkowników. Dodatkowym wynikiem tej analizy są reguły *IF-THEN*, które używane są do tworzenia sugestii dla użytkowników. Badania dotyczące systemów rekomendacyjnych polegają na adaptacji mechanizmów wspomagania decyzji w procesach decyzyjnych. Proponuję tutaj zastosowanie porządku leksykograficznego. Pozwala on na symulację ludzkiego procesu określania wartości wielu alternatyw oraz ich uporządkowanie.

Zakres moich badań uwzględnia również zastosowania technik inteligencji obliczeniowej do **konstrukcji algorytmów do wzbogacania informacji w oparciu o ontologie**. Celem tych badań jest

budowa systemów wyposażonych w mechnizmy ekstrakcji informacji na poziomie konceptów i pojęć. Badania te skupiają się na wzbogaceniu i wykorzystaniu hierarchii konceptów (ang. hierarchy of concepts). Algorytmy do automatycznego określania ważności konceptów zostały zaprezentowane i wykorzystane do ulepszenia hierarchii konceptów. Wzbogacone o elementy z ontologii hierarchie umożliwiają identyfikację konceptów w dokumentach sieciowych.

Ważnym kierunkiem moich badań jest zastosowanie inteligencji obliczeniowej do **przetwarzania danych przedstawionych w formacie RDF** – Resource Description Framework. Rezultatem tych badań są methody do oszacowania podobieństwa pomiędzy dwoma opisami dowolnych zasobów sieciowych. Zasobami tymi mogą być jakiekolwiek informacje przechowywane na webie. Metody te stosowane są w metodologii do przeszukiwania rozproszonych danych sieciowych, reprezentowanych w różnych formatach danych. Proponowana metoda określania podobieństwa stanowi podstawę do konstrukcji algorytmów do automatycznej budowy definicji kategorii w oparciu o dostępne dane.

IV.3 Szczegółowy opis problemów badawczych

IV.3.1 Analiza danych z portali społecznościowych

Coraz częściej web postrzegany jest jako forum społeczne – to znaczy, że użytkownicy traktują web jako miejsce interakcji socjalnych, miejsce nawiązania kontaktu z innymi użytkownikami lub grupami o podobnych zainteresowaniach [L1][4][L2]. Zachowanie takie obserwujemy w portalach społeczościowych. Niektóre z tych portali wyposażone są w mechanizmy tak zwanego 'znakowania' (ang. tagging) [L3][L4][L5]. Oznacza to, że użytkownik znakuje (oznacza etykietą) przeglądany przez siebie zasób. W ten sposób wyraża on swoją opinię dotyczącą tego zasobu – może go opisać, wyrazić pogląd co o nim myśli lub co chciałby z nim zrobić. W ogólności, pojedyńczy użytkownik może oznakować wiele zasobów używając wielu etykiet. Jako rezultat, oznakowane zasoby i użyte etykiety charakteryzują użytkownika [L6].

W pracy [R3] proponuję metodę, która reprezentuje użyte etykiety i oznakowane przez pojedyńczego użytkownika zasoby w postaci zbiorów rozmytych (ang. fuzzy sets) [L7]. Zbiór rozmyty przedstawiający etykiety użyte przez użytkownika do oznakowania zasobów wyglada natepująco:

$$TagPop_u(t) = \{\frac{a_1}{t_1}, \frac{a_2}{t_2}, \ldots, \frac{a_n}{t_n}, \ldots\} \tag{1}$$

gdzie

$$a_n = \frac{number\ of\ times\ t_n\ is\ used\ by\ the\ user}{max\ number\ of\ resources\ taged\ by\ the\ user\ with\ t_n} \tag{2}$$

oraz t_n przedstawia użyte etykiety. Powyższy zbiór rozmyty reprezentuje popularność etykiet.

Inny zbiór rozmyty jest utworzony na podstawie wszystkich zasobów, które użytkownik oznakował. Zbiór taki reprezentuje ważność zasobów dla tego użytkownika:

$$ResAttract_u(r) = \{\frac{b_1}{r_1}, \frac{b_2}{r_2}, \ldots, \frac{b_m}{r_m}, \ldots\} \tag{3}$$

gdzie

$$b_m = \frac{number\ of\ tags\ used\ for\ r_m\ by\ the\ user}{max\ number\ of\ tags\ used\ for\ a\ single\ resource\ by\ the\ user} \tag{4}$$

[4]Przyjęta konwencja cytacji: R – publikacje cyklu, L – ogólne publikacje (sekcja: Literatura), c – publikacje konferencyjne (załącznik: Całościowy Wykaz → Publikacje konferencyjne recenzowane).

a r_m oznacza pojeczyńczy zasób. Obydwa zbiory przedstawiają ważność połączeń pomiędzy etykietami i zasobami. Jeśli obydwa zbiory rozmyte utworzone z danych o tym samym użytkowniku zostaną połączone w relację, to otrzymamy sygnaturę użytkownika:

$$UserSignature_u(r,t) = ResAttrac_u(r) \times TagPop_u(t) \tag{5}$$

Tak zdefiniowane sygnatury mogą być użyte do oszacowania podobieństwa pomiędzy użytkownikami. W tym celu zaproponowana została miara podobieństwa pomiędzy użytkownikami nazwana 'pokrewieństwo' (ang. kindredness). Polega ona na porównaniu sygnatur i oparta jest na indeksie Jaccard'a. Porównanie wymaga określenia liczb kardynalnych zbiorów rozmytych. Proponuję tu specjalny proces określania liczb kardynalnych, w którym użytkownik podaje jaka wartość stopnia przynależności wymagana jest aby dany element należał do zbioru. Innymi słowy, wartość ta reprezentuje percepcję użytkownika odnośnie 'spójności' zbioru. Fakt, że każdy użytkownik może podać inną wartość, oznacza, że możemy otrzymać różne liczby kardynalne. Prowadzi to do personalizacji procesu oszacowania podobieństwa pomiędzy użytkownikami.

Ostatecznie, miara 'pokrewieństwa' użytkowników wyrażona jest w następujący sposób:

$$interestBasedComp(u_i,u_j) = \frac{|T(UserSignature_{u_i}(r,t), UserSignature_{u_j}(r,t))|}{|UserSignature_{u_i}(r,t)|} \tag{6}$$

gdzie | | oznacza liczbe kardynalną zbioru.

Jednym z rezultatów badań jest również metodologia agregacji sygnatur użytkowników, którzy są członkami tej samej grupy. Otrzymana sygnatura reprezetuje całą grupę, jej wszystkich członków:

$$GroupSignature_u(r,t) = AGR_{u_i \in G}(UserSignature_{u_i}(r,t)) \tag{7}$$

Rozważanym operatorem agregacji AGR jest powszechnie stosowany operator OWA (Ordered Weighted Aggregator) [L8][L9]. Wybór ten podyktowany jest zdolnością OWA do agregacji informacji używając kwantyfikatorów lingwistycznych (ang. linguistic quantifiers). Kwantyfikatory te pozwalają na kontrolę stopnia udziału pojeczyńczych informacji w 'budowie' zagregowanej wartości. Użycie różnych kwantifikatorów prowadzi do uwzględnienia różnych informacji w trakcie tworzenia sygnatury. Przykładowo, jeśli użyjemy kwantyfikator 'MAX', utworzony GroupSignature będzie zawierał etykiety i zasoby popularne nawet pomiędzy kilkoma użytkownikami. Tak utworzona sygnatura reprezentować będzie pełen zakres etykiet i zasobów włączając te, które używane były bardzo rzadko. W przypadku kwantyfikatora 'MIN', utworzona sygnatura zawierać będzie tylko najczęściej używane etykiety i najbardziej popularne zasoby pomiędzy członkami grupy.

Sygnatura grupy wykorzystywana jest do porównania grupy z pojedyńczym użytkownikiem.

$$interestBasedCompGroup(u_i,G) = \frac{|T(UserSignature_{u_i}(r,t), GroupSignature(r,t))|}{|UserSignature_{u_i}(r,t)|} \tag{8}$$

W ten sposób, użytkownik może wyszukiwać grupy, których członkowie mają podobne zainteresowania. Dodatkowo, użytkownik ma kontrolę nad procesem porównania użytkownika z grupą. Jest to możliwe dzięki przedstawionym wcześniej kombinacjom wymagań dotyczących obliczeń liczb kardynalnych zbiorów rozmytych, jak również budowy sygnatury grupy. Możemy wyróżnić następujące przypadki:

- Porównanie 'swobodne': użytkownik zamierza znaleźć nowe zasoby. Porównanie odbywa się ze wszystkimi członkami grupy. Użytkownik nie spodziewa się, że wszyscy będą podzielać jego zainteresowania; nawet pojedyńczy użytkownik posiadający jego zainteresowania wystarczy, aby zaintersować się tą grupą.

- Porównanie 'skoncentrowane': użytkownik jest bardzo precyzyjny w swoich poszukiwaniach i oczekuje, że wszyscy członkowie grupy bedą posiadać takie same, zgodne z jego, zainteresowania. Dodatkowo grupa powinna być jednolita w przypadku używanych etykiet i popularnych zasobów.
- Porównanie 'większościowe': użytkownik jest umiarkowany w poszukiwaniach nowych zasobów, ale jednocześnie wymaga on, aby zdecydowana większość użytkowników podzielała jego zainteresowania.

Przytoczone przykłady odzwierciedlają elastyczność i możliwość kontroli proponowanego procesu porównawczego.

Potwierdzeniem atrakcyjności zastowawania zbiorów rozmytych w portalach społecznościowych są eksperymenty przedstawione w publikacjach. Praca [R3] zawiera realistyczny przykład zastosowania proponowanej metody. Dane zgromadzone z portala www.librarything.com reprezentują informacje dotyczące kilkunastu użytkownikow. Zawierają one detale dotyczące książek oraz etykiet, które użytkownicy wykorzystali do oznakowania tych książek.

IV.3.2 Budowa profili użytkowników i systemów rekomendacyjnych

Celem licznych działań badawczych związanych z siecią internetową jest przekształcenie sieci w przyjazne dla użytkowników środowisko, gdzie mogą oni łatwo znaleźć informację, której szukają. Oznacza to, że proces szukania użytecznej informacji powinien być szybki i skuteczny. Jednakże rosnący zasób dostępnej informacji powoduje, że stworzenie systemów wspomagających użytkowników w wyszukiwaniu odpowiedniej informacji nie jest zadaniem łatwym. W wielu przypadkach systemy takie wymagają profili użytkowników. Profile reprezentują zainteresowania użytkowników i są używane do identyfikowania istotnych dla nich informacji [L10][L11][L12].

Proces uaktualnienia profilu użytkownika jest przedmiotem badań przedstawionych w pracach [R8][c34]. W artykułach proponuję nową metodę automatycznej aktualizacji profilu użytkownika. Metoda ta oparta jest na proponowanej mierze trafności (ang. relevance measure), która jest kombinacją dwóch miar: 1) podobieństwa pomiędzy elementami stron portalowych odwiedzanych przez użytkownia a elementami z jego profilu; oraz 2) znaczenia (ważności) elementów ze stron portalowych. Obydwie miary sa równoważne przy ustalaniu miary trafności.

Proponowana metoda opiera się na analizie aktywności użytkownika i identyfikacji najistotniejszych elementów i faktów, które powinny być dodawane do jego profilu. Proces wyboru elementów i faktów wykorzystuje semantyczną miarę podobieństwa (ang. semantic-based similarity measure). Wartość tej miary oszacowywana jest poprzez użycie reguł reprezentujących różne aspekty podobieństwa. Reguły te skonstruowane są w oparciu o ontologie domen reprezentujących zainteresowania użytkownika. Wartości otrzymane w wyniku aktywowania (odpalenia) reguł podlegają agregacji. W tym przypadku używany jest operator OWA (ang. Ordered Weighted Aggregating) [L8]. Operator ten pozwala nam kontrolować proces agregacji z punktu widzenia sposobu łączenia agregowanych wartości. Kontrola ta odbywa się poprzez kwantyfikatory lingwistyczne takie jak 'LUB' (ang. 'OR'), 'KILKA' (ang. 'SOME'), czy 'WIĘKSZOŚĆ' (ang. 'MOST'). Otrzymane wartości miar podobieństwa są następnie łączone z miarami ważności poszczególnych elementów. Ważność elementów oszacowywana jest na podstawie historii przeglądania stron portalowych przez użytkownika. W ten sposób identyfikowane są, i dołączone do profilu użytkownika, najbardziej dopowiednie dla niego elementy. Proponowana metoda została zastowowana do aktualizacji profili użytkowników zainteresowanych w szukaniu i słuchaniu muzyki. W aplikacji wykorzystane zostały dane z portali MusicBrainz i Wikipidia.

Niezaprzeczalna ważność profili użytkowników skłoniła mnie do dalszych badań dotyczących metod i technik wspomagających użytkowników w procesach szukania na sieci odpowiednich dla nich danych i usług.

Typowym przykładem implementacji takich metod i technik są systemy rekomendacyjne [L13][L14][L15]. Stanowią one ważny element internetu. Ich obecność i użycie prowadzą do lepszego i bardziej przyjaznego dla użytkowników korzystania z sieci. Systemy rekomendacyjne powinny być wyposażone w mechnizmy analizy recenzji dostarczonych przez użytkowników [L16][L17][L18]. Przedstawione wyniki badań są rezultatem prac nad konstukcją metod posiadających możliwości 'naśladowania' mechanizmów wyboru charakterystycznych dla człowieka.

Specjalna metoda selekcji/wyboru wykorzystująca preferencje podane przez użytkownika przedstawiona jest w pracach [R6][R10][c38]. Istotnym jej elementem jest leksykograficzne uporządkowanie preferencji. Metoda jest prosta i zapewnia efektywną selekcję najbardziej odpowiedniej alternatywy.

Proponowana metoda łączy informacje dotyczące uporządkowanych kryteriów ze stopniami spełnienia tych kryteriów. W pracy przedstawiam i opisuję algorytm oceniający każdą z alternatyw w oparciu o zestaw kryteriów określonych przez użytkownika. Algorytm wykorzystuje proces agregacji oparty o srednią ważoną i dynamicznie oblicza wagi w oparciu o leksykograficzne uporządkowanie kryteriów. Kryteria uporządkowane są w oparciu o ich priorytety. Wartości wag obliczane są individualnie dla każdej alternatywy – zależą one od lokalizacji danego kryterium w hierarchii kryteriów oraz od stopnia satysfakcji kryterium o wyższym priorytecie.

Ogólna ocena alternatywy obliczana jest w oparciu o poniższe równanie:

$$Score(x_k) = \sum_{i=1}^{n} w_i C_i(x_{k,(i)}) \tag{9}$$

gdzie $C_i(x_{k,(i)})$ jest wartością reprezentującą ocenę (stopień satysfakcji) kryterium C_i dla atrybutu i alternatywy x_k (każde kryterium C_i jest 'związane' z atrybutem i). Wagi w_i pozwalają nam kontrolować ważność każdego z kryteriów i jego wkład w ogólną ocenę alternatywy.

Rozważmy sytuację, w której mamy do czynienia z leksykograficznie uporządkowaną listą kryteriów podaną przez użytkownika. Zakładamy tutaj uprządkowanie liniowe, w którym dwa kryteria nie mają takiego samego priorytetu. W tym przypadku [L9], waga w_i związana z kryterium C_i zależna jest od stopnia satysfakcji poprzedniego (w rankingu) kryterium. Koncept ten jest bezpośrednio odzwierciedlony w metodzie obliczania wartości wagi w_i. Oznacza to, że wartość wagi w_i zależy od stopnia satysfakcji kryteriów: C_1, C_2, ..., C_{i-1}.

Proces obliczania wag jest następujący. Zakładamy, że u_1, u_1, ... u_n są wagami początkowymi. Ich wartości obliczane są następująco:

$$\begin{aligned} u_1 &= 1 \\ u_2 &= C_1(x_{k,(1)}) \\ u_3 &= C_1(x_{k,(1)}) \times C_2(x_{k,(2)}) = u_2 \times C_2(x_{k,(2)}) \\ &\dots \\ u_n &= C_{n-1}(x_{k,(n-1)}) \end{aligned}$$

Ważne jest aby ogólna ocena alterantywy, której atrybuty maksymalnie spełniają wszysktie kryteria była równa 1.0. Aby to uzyskać, wartości wag początkowych u_i są normalizowane. Rezultatem są wagi w_i:

$$w_i = \frac{u_i}{n} \tag{10}$$

Ocena satysfakcji poszczególnych kryteriów odbywa się wykorzystując prostą funkcję satysfakcji. Konstrukcja tęj funkcji wymaga podania przez użytkownika wartości pojedyńczego parametru, który reprezentuje

granicę pomiędzy wartościami akceptowalnymi dla niego a wartościami nieakceptowalnymi. Oznacza to, że stopień satysfakcji $C_i(x_{k,(i)})$ otrzymywany jest jako wartość funkcji satysfakcji obliczonej dla wartości atrybutu i.

Zaproponowana metoda obliczania stopnia satysfakcji alternatywy została zaimplementowana i stanowi podstawę aplikacji webowej PET (ang. Personal Evaluation Tool). Celem aplikacji jest wspomaganie użytkownika w dokonaniu wyboru właściwej alternatywy. Aplikacja posiada możliwość wspomożenia użytkownika w wyborze muzyki – Serwer Pobierania Muzyki, i w wyborze hotelu – Serwer Rekomendacji Hoteli. Przeprowadzone zostały kompleksowe badania PET's. Wyniki uzyskane z PET'a pozostały w dużej korelacji z wyborami wykonanymi przez użytkowników. Jedną z ciekawych obserwacji jest widoczny dowód zmiennej natury użytkownika, to znaczy, że użytkownik potrafi zmienić swoje preferencje.

Opisana wcześniej metoda budowy zbiorów rozmytych w oparciu o użyte etykiety i zasoby, została połączona z leksykograficznym porządkowaniem preferencji w pracy [R7]. Jako rezultat otrzymaliśmy nowe podejście do ustalenia najbardziej odpowiednich rekomendacji. Główna idea polega na przedstawieniu opinii użytkowników jako zbiory rozmyte.

Dalsze prace badawcze w tym zakresie uwzględniają zastosowanie zbiorów rozmytych do konstrukcji listy przedmiotow, które sugerowane są użytkownikowi [c56] oraz budowy systemów przyjaznych dla użytkownika (ang. user-friendly), które 'działają' jak człowiek (ang. human-like), gdy proces wymaga selekcji jednej z wielu alternatyw [c57].

Badania dotyczące interakcji użytkownika z systemami z punktu widzenia zapytań i odpowiedzi rozszerzone zostały na tematykę 'obliczeń na słowach' (ang. computing with words – CW) [L19]. Połączenie 'obliczeń na słowach' z ontologią jest przedmiotem publikacji [R12]. Badania koncentrują się na stworzeniu systemu, który daje użytkownikowi możliwość interakcji słownej (ang. lingustic interface). Ważnymi elementami tego systemu są ontologie i systemy wnioskowania przybliżonego. Zastowoanie ontologii zapewnia dużą ekspresywność, możliwość użycia różnych typów danych oraz personalizację. W szczególności:

- ontologia wzbogaca semantykę predykatów występujących w CW: każdy termin predykatu (ang. predicate term) jest instancją konceptu zdefiniowanego w ontologii, co oznacza, że semantyka terminu jest w pełni zdefiniowana ontologicznie; definicja taka dostarcza dodatkowych informacji o terminie w formie relacji tego terminu z innymi terminami zdefiniowanymi w ontologii;
- ontologia umożliwia prosty sposób personalizacji ograniczeń, które można nałożyć na definicje terminów i relacji; ograniczenia te mogą być narzucone na definicje przez pojedyńczych użytkowników lub grupy użytkowników;
- ontologia zapewnia uniwersalną reprezentację predykatów i reguł; języki reprezentacji ontologii takie jak RDF i OWL, które zbudowane są w oparciu o XML, dostarczają jednakowy format przedstawiania informacji; oznacza to, że każdy predykat i reguła mogą być przetwarzane przez jakiekolwiek systemy wnioskowania.

Przedstawione zalety ontologii prowadzą do stworzenia środowiska, które umożliwia konstrukcję predykatów o bogatej treści. Takie środowisko wzbogaca procesy obliczeń na słowach poprzez pełniejsze wykorzystanie relacji semantycznych i systemów wnioskowania.

IV.3.3 Konstrukcja algorytmów do wzbogacania informacji w oparciu o ontologie

Badania nad postrzeganiem dokumentów jako kolekcji konceptów a nie zbioru słów są tematem następnego etapu moich badań naukowych. Z punktu widzenia wyszukiwania informacji, traktowanie dokumentów we-

bowych jako kolekcji konceptów prowadzi do innego spojrzenia na proces identyfikacji dokumentów ważnych dla użytkownika. Tekst będący kolekcją słów, jest w rzeczywistości zbiorem definicji konceptów. W związku z tym, aby 'znaleźć' w dokumentach koncepcję, wymagana jest znajomość ich definicji i możliwość odszukania ich w tekscie [L20][L21][L22][L23][L24].

W pracy [R9] proponuję nową metodę określania ważności pojedyńczych słów. Metoda ta nosi nazwę AATI (ang. adaptive assignment of term importance). Jest ona oparta o ontologię, która używana jest jako zbiór definicji konceptów i służy do ich identyfikacji. Metoda ta zawiera następujące komponenty: definicje relacji pomiędzy słowami i konceptami oraz iteracyjny algorytm do określania ważności słów.

AATI ustawicznie uaktualnia ważność słów w oparciu o ich występowanie w dokumentach. Metoda ta oparta jest na dwóch miarach: ważność słowa (ang. term weight – TW) i wartość strony (ang. page value – PV). Iteracyjny algorytm oblicza wartości TW w oparciu o dokumenty. AATI charakteryzuje się następującymi cechami:

- słowa, które stanowią definicje konceptów są częścią ontologii danej dziedziny zainteresowań;
- wartości TW obliczane są na bieżąco według następującej idei: jeśli słowo znajduje się w dokumencie, który zawiera związany z nim koncept, wartość TW tego słowa rośnie, jeśli natomiast dane słowo pojawia się w dokumentach rzadko, jego TW wartość maleje; oznacza to, że proces obliczania TW zależy od zbioru przeglądanych dokumentów webowych;
- wartości TW są uaktualniane w tym samym czasie kiedy PV dokumentów oszacowywane jest bez jakiejkolwiek wiedzy o istnieniu konceptów w dokumencie; proces ten nie wymaga fazy treningu, który byłby konieczny w przypadku konstrukcji klasyfikatora;
- wartości TW obliczane są w oparciu o wartości PV, a wartości PV obliczane są w opariu o TW; początkowe wartości TW generowane są losowo.

Przedstawione powyżej cechy powodują, że AATI posiada zdolności dostosowania się do zmian i do radzenia sobie z różnymi dokumentami. Oznacza to, że AATI jest przydatny do 'dostrajania' definicji konceptów. Definicje konceptów znajdujące się w ontologii zbudowane są ze słów. Jednakże, nie wszystkie te słowa są równo ważne z punktu widzenia definicji danego konceptu. AATI pozwala nam na zidentyfikowanie stopnia ważności tych słów poprzez obliczanie ich wartości TW.

AATI został zastosowany do konstrukcji inteligentnego systemu do identyfikacji dokumentów zawierających koncepty, które odpowiadają zestawowi słów kluczowych podanych przez użytkownika [R5][R11]. Proponowana metodologia modeluje strukturę konceptów w oparciu o dostarczone słowa kluczowe. Zbudowany model przedstawiony jest jako hierarchia konceptów (ang. hierarchy of concepts – HofC) [L25]. Skonstruowana hierarchia wzbogacona jest poprzez słowa i definicje pokrewnych konceptów znajdujących się w ontologii.

Po skonstruowaniu i wzbogaceniu przez dodanie informacji z ontologii, HofC porównany jest z dokumentem. Porównanie takie oznacza sprawdzenie ile i jakie słowa znalezione w dokumencie wchodzą w skład HofC. Każde słowo znalezione w dokumencie 'aktywuje' odpowiedni węzeł w HofC. Poziomy aktywacji poszczególnych węzłów podlegają agregacji.

Identyfikacja konceptu w dokumencie jest równoznaczna z określeniem poziomu aktywacji części HofC reprezentującej ten koncept. Im więcej węzłów HofC jest aktywowanych tym bardziej aktywowany jest cały HofC. Jak wspomniano wcześniej, cały HofC przestawia zestaw słów kluczowych reprezentujących szukane przez użytkownika informacje. Im bardziej aktywowany jest HofC poprzez słowa z dokumentu tym bardziej dany dokument odpowiada szukanej informacji.

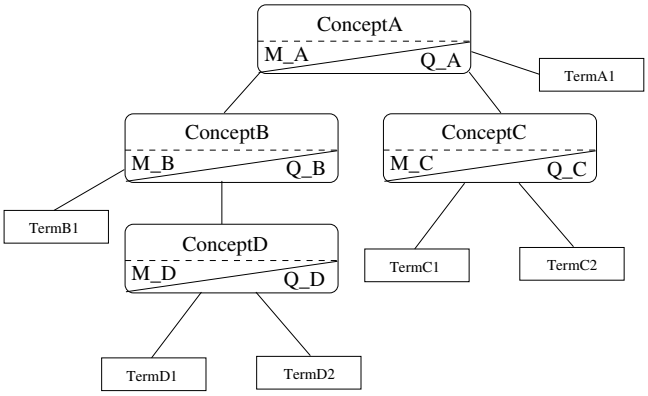

Figure 1: Wzbogacona hierarchia konceptów

Rysunek 1 przedstawia prosty HofC. Zawiera on centralny koncept A, którego definicja składa się z dwów innych konceptów B i C, oraz pojedyncze słowo/termin A1. Koncept B definiowany jest przez koncept D oraz słowo/termin B1. W przypadku konceptów C i D, są one zdefiniowane przez słowa/terminy C1 i C2, oraz D1 i D2.

Jak widzimy na rysunku, węzły z konceptami – A, B, C i D – posiadają dwa symbole: $M\_$ oraz $Q\_$. Pierwszy z nich reprezentuje wektor o długości równej liczbie krawędzi wychodzących z węzła. Każdy element wektora opowiada ważności odpowiedniego podwęzła w definicji węzła. Jest to wektor ważności. Drugi z symboli reprezentuje kwantyfikator lingwistyczny kontrolujący proces agregacji stopni aktywacji słów i konceptów związanych z podwłęzami. Zarówno $M\_$ jak i $Q\_$ użyte są do obliczenia wag operatora OWA.

Obliczenie stopnia aktywacji pojedyńczego słowa/terminu odbywa się w oparciu o równanie:

$$S_i = e^{-\frac{1}{freq_i}} \tag{11}$$

gdzie $freq_i$ reprezentuje częstotliwość występowania słowa w dokumencie. Przedstawione równanie posiada dwie istotne cechy: 1) jest to monotonicznie rosnąca funkcja, co oznacza, że poziom aktywności wzrasta z częstotliwością występowania słowa; 2) funkcja ta nasyca się; zapobiega to nadmiernej dominacji danego słowa; jeśli dane słowo występuje w dokumencie wielokrotnie, poziom jego aktywności nie rośnie w nieskończoność. Po obliczeniu stopnia aktywacji każdego słowa/terminu, algorytm oblicza stopnie aktywacji konceptów, które zdefiniowane są wyłącznie przez słowa/terminy (rysunek 1), koncepty D i C (rysunek 1). Stopień aktywacji konceptu wynosi:

$$S_c = \sum_{k=1}^{n} w_k * S_k^{uporzadkowane} \tag{12}$$

gdzie n oznacza liczbę słów/terminów definiujących koncept, $S_k^{uporzadkowane}$ oznacza uporządkowane od największego do najmniejczego poziomy aktywacji słów, a wagi w_k obliczone są wykorzystując $M\_$ oraz $Q\_$. Dla przykładu, stopień aktywacji konceptu D obliczany jest w oparciu o stopnie aktywacji słów/terminów D1 i D2.

W przypadku, gdy koncept zdefiniowany jest przez inne koncepty i słowa (koncepty A i B na rysunku), poziom ich aktywacji obliczany jest następująco: jako pierwsze agregowane są stopnie aktywności konceptów,

następnie agregacji podlegają stopnie aktywacji słów. Ostatnim etapem jest agregacja obydwu obliczonych wartości.

Stopień aktywacji konceptu A odbywa się w dwóch fazach: 1) stopień aktywacji słowa/terminu A1 obliczany jest na postawie częstotliwości jego wystąpienia; 2) stopnie aktywacji konceptów B i C obliczane są w oparciu o słowa/terminy i koncepty je definiujące. Ostatecznie wszystkie stopnie aktywacji, w tym przypadku trzech słów/terminów i dwuch konceptów, agregowane są poprzez użycie agregatora OWA z wagami obliczonymi na podstawie wartości wektorów M_A i Q_A.

Podsumowywując, aktywacja słów i konceptów HofC jest propagowana w górę. Oznacza to, ze obliczenia aktywacji zaczynają się od słów na najniższym poziomie (dnie) grafu HofC, a kończą się na najwyższym. Obecność słów D1, D1, C1 oraz C2 przyczynia się do aktywacji konceptów D i C. Aktywacja konceptu D połączona z aktywacją słowa B1 użyta jest do aktywacji konceptu B. Następnie aktywacja konceptów B i C oraz słowa A1 prowadzi do aktywacji konceptu A.

Wspomniany powyżej kwantyfikator lingwistyczny używany do określenia wag operatora OWA zdeterminowany jest przez użytkownika i może być inny dla każdego węzła grafu HofC. Możliwe kwantyfikatory to: 'WIĘKSZOŚĆ' (ang. 'MOST'), 'CO NAJMNIEJ POŁOWA' (ang. 'A LEAST HALF'), 'WSZYSTKO' (ang. 'ALL') i 'OKOŁO JEDNEJ TRZECIEJ' (ang. ABOUT ONE THIRD).

Prace [R5][R9][R11]zawierają kompleksowe przykłady użycia proponowanej metody. Otrzymane rezultaty i rewizja procesu obliczeniowego potwierdziły możliwość zastosowania metody do identyfikacji konceptów w dokumentach. Rezultaty identyfikacji były bardzo podobne do identyfikacji przeprowadzonej przez użytkowników. Potwierdza to, że synteza wielu technik i informacji: ontologii, AATI, HofC, kwantyfikatorów lingwistycznych oraz operatora agregacji OWA, prowadzi do konstrukcji systemów o zwiększonych i podobnych do ludzkich możliwościach.

IV.3.4 Przetwarzania danych przedstawionych w fomacie RDF

Tekst i dokumenty są najczęstszą formą informacji dostępnej na sieci. Składowana tak informacja trudna jest do kompleksowej analizy i przetwarzania. Bardzo atrakcyjną i stosunkowo nową formą reprezentacji danych i informacji stanowi RDF (ang. Resource Description Framework (RDF) [L26]. RDF reprezentuje reguły przedstawiania i składowania danych, jak również zawiera reguły definiowania metadanych (ang. metadata).

RDF definiuje bardzo atrakcyjną formę reprezentacji danych, która oparta jest o bardzo prostą regułę przedstawiania informacji w formie trójki $< subject - property - object >$. Oznacza to, że encje (ang. entities) połączone są relacją, która istnieje pomiędzy tymi dwoma encjami. Na przyklad, trójka RDF $< Jan - mieszkaW - Warszawa >$ łączy *Jana* z miatem *Warszawą* poprzez ralację *mieszkaW*. Oznacza to, że istnieje zależność pomiędzy Janem i Warszawą. Inne trójki, na przykład $< Jan - jest - osoba >$ i $< Warszawa - jest - miasto >$ dostarczają nowych informacji. Jak widzimy, wszystkie te informacje są ze sobą połączone. Ta naturalna cecha łączenia informacji jest bardzo istotną zaletą RDF. RDF staje się bardzo popularną formą reprezentacji informacji w sieci internetowej, jak również w aplikacjach przemysłowych [L27]. Jego atrakcyjność spotęgowana jest nowymi możliwościami analizy i przetwarzania.

Jedną z bardziej powszechnych czynności, która jest częścią wielu metod i algorytmów działających na danych, jest określenie podobieństwa pomiędzy dwoma encjami informacji [L28][L29]. Kluczowym elementem przedstawionej przeze mnie czwartej tematyki badawczej jest metodologia obliczania podobieństwa danych w formacie RDF, którą nazywamy podobieństwem 'opartym-na-cechach' (ang. feature-based similarity).

W pracach [R4][R2] proponuję i opisuję nową metodę obliczeń podobieństwa semantycznego danych w formacie RDF. Metoda ta oparta jest na zależnościach pomiędzy elementami trójek RDF. Dodatkowo,

metoda ta wykorzystuje elementy teorii możliwości (ang. possibility theory), to znaczy, miarę konieczności (ang. necessity) i miarę możliwości (ang. possiblity). Określone w ten sposób podobieństwo pozwala na oszacowanie dolnej i górnej granicy podobieństwa pomiędzy jakąkolwiek parą trójek RDF.

Ważną cechą proponowanej metody jest możliwość określenia podobieństwa w sytuacjach kontekstowych. Jest to możliwe poprzez określenie podobieństwa tylko w oparciu o ściśle określone relacje odpowiednie dla danego kontekstu. Inną cechą metody, która może wywoływać dyskusję, jest asymetria. Jest to zgodne z poglądami Tversky'go [L30] i Nosofsky'go [L31], którzy gloszą, że kierunek asymetrii podobieństwa jest odzwierciedleniem różnorodności znaczenia porównywanych encji. Sytuacja taka może zaistnieć w przypadku opisu encji za pomocą trójek RDF jeśli jedna z encji posiada wiele cech, a druga ma ich niewiele. Niesymetria oznacza, że jedna z encji jest mniej znacząca i jednocześnie bardziej podobna do encji, która jest bardziej znacząca. Bardziej znacząca encja jest jednocześnie mniej podobna do encji mniej znaczącej. Taka ocena podobieństwa może być zastosowana w przypadku potrzeby określenia znaczenia porównywanych encji.

Popularność RDF i jego użycie jako formy reprezentacji danych doprowadziło do powstania interesującej formy internetu, który nazywa się LOD (Linked Open Data) [L32]. Głównym formatem używanym do reprezentacji danych jest RDF. LOD jest siatką połączonych zasobów/encji reprezentowanych jako trójki RDF. Formalnie

$$LD = \{< r_i, p_q, r_m >: r_i, r_m \in R, p_q \in P\} \tag{13}$$

gdzie R jest zbiorem zasobów a P zbiorem relacji. Pojedyńczy zasób r_i zdefiniowany jest poprzez jego połączenia z innymi zasobami. Każdy z tych zasobów uważany jest za cechę zasobu r_i. Zbiór zasobów połączonych do r_i traktowany jest jako jego definicja semantyczna. Połączenia pomiędzy zasobem r_i i innymi zasobami oznakowane są relacjami. Dlatego też dla danego zasobu r_i możemy napisać:

$$n^i = |\{< r_i, p_q, r_m >: r_m \in R \setminus r_i, p_q \in P\} \tag{14}$$

gdzie $|\ |$ oznacza liczbę kardynalną zbioru, a n^i reprezentuje liczbę połączeń pomiędzy r_i a innymi zasobami połączonymi do r_i. Innymi słowy, n^i reprezentuje liczbę cech zasobu r_i. Taka interpretacja oznacza, że LOD jest środowiskiem definiującym semantykę zasobów. LOD postrzegana jest jako semantyczna przestrzeń zawierająca definicje, które są 'wymieszane i połączone' pomiędzy sobą. Dlatego też możemy wykorzystać LOD do obliczania semantycznego podobieństwa traktując połączenia jako wskaźniki 'pokrewieństwa' pomiędzy zasobami. Aby określić podobieństwo dwóch zasobów r_i i r_j identyfikujemy i analizujemy ich cechy.

Analiza cech dwóch zasobów/encji prowadzi do zdefiniowania czterech przypadków podobieństwa pomiędzy nimi. Na przykład, niektóre zasoby nie są współdzielone pomiędzy rozpatrywanymi zasobami, w innym przypadku takie same zasoby mogą być połączone innymi relacjami. Każdy z czterech przypadków reprezentuje wkład do podobieństwa lub odmienności (ang. dissimilarity). Detale opisujące wszystkie cztery przypadki przedstawione są poniżej:

- S1: przypadek ten stanowi niezaprzeczalny wkład do podobieństwa; takie same relacje łączą współdzielony zasób z obydwoma porównywanymi zasobami.
- S2: przypadek ten wprowadza niejasność (ang. ambiguity); takie same relacje łączą obydwa porównywane zasoby z innymi zasobami.
- S3: również i ten przypadek wprowadza niejasność: współdzielony zasób połączony jest z porównywanym zasobem poprzez inne relacje. W tym przypadku, w zależności od pokrewieństwa/podobieństwa relacji przypadek ten może przedstawiać wkład do podobieństwa lub odmienności.
- S4: Ostatni, czwarty przypadek przedstawia wkład do odmienności. Każdy z porównywanych zasobów połączony jest z innym zasobem inną relacją.

Formalnie, przedstawione powyżej przypadki wyrażone są w następujący sposób: jeśli r_i oznacza zasób, wtedy liczba egzemplarzy S1 wynosi:

$$n^i S1 = |\{(< r_i, p_q, r_m >, < r_j, p_q, r_m >) : r_m \in R_i \cap R_j, p_q \in P_i \cap P_j\}| \tag{15}$$

liczba egzemplarzy S2:

$$n^i S2 = |\{< r_i, p_q, r_m >: r_m \in R_i \setminus R_j, p_q \in P_i \cap P_j\}| \tag{16}$$

liczba egzemplarzy S3:

$$n^i S3 = |\{(< r_i, p_q, r_m >, < r_j, p_s, r_m >) : r_m \in R_i \cap R_j, (p_q \neq p_s) \in P\}| \tag{17}$$

oraz liczba egzemplarzy S4:

$$n^i S4 = |\{< r_i, p_q, r_m >: r_m \in R_i \setminus R_j, p_q \in P_i \setminus P_j\}| \tag{18}$$

W oparciu o przedstawione liczby egzemplarzy każdego przypadku, podobieństwo i odmienność zasobów r_i i r_j mogą być określone. Podobieństwo zależy wyłącznie od przypadku S1. W związku z tym miara konieczności [L33][L34]podobieństwa wynosi:

$$N(sim[r_i, r_j]) = \frac{n^i S1}{n^i} \tag{19}$$

Ponieważ równanie to przedstawia podobieństwo r_j do r_i, mianownik zawiera ogólna liczbę cech zasobu r_i. Miara konieczności odmienności określona jest przez przypadek S4, tak więc równanie ma postać:

$$N(dissim[r_i, r_j]) = \frac{n^i S4}{n^i} \tag{20}$$

Jak wspomniałem wcześniej, przypadki S2 i S3 wprowadzają niejasność. Wykorzystujemy je do oszacowania miary możliwości dotyczącej odmienności:

$$\Pi(dissim[r_i, r_j]) = \frac{n^i S2 + n^i S3 + n^i S4}{n^i} \tag{21}$$

uznając, że miara konieczności odmienności jest częścią miary możliwości odmienności. Jest to zgodne z zasadą minimalnej specyfiki (ang. minimal specificity), że jeśli nie wiadomo czy dany egzemplarz jest niemożliwy, to jest on uznany za możliwy [L35].

Ostatecznie, jeśli weźniemy pod uwagę fakt, że n^i jest równe sumie egzemplarzy wszystkich przypadków $\sum_{k=1,2,3,4} n^i Sk$, prowadzi to do następujących wartości:

$$N(sim[r_i, r_j]) = \frac{n^i S1}{n^i} \tag{22}$$

$$\Pi(sim[r_i, r_j]) = 1 - \frac{n^i S4}{n^i} \tag{23}$$

Dlatego też podobieństwo pomiędzy zasobami może być wyrażone jako przedział pomiędzy $N(sim)$ a $\Pi(sim)$. Jest to ważny aspekt proponowanej metody. Dolna granica przedziału oznacza wartość podobieństwa związaną z dużym poziomem pewności. Górna granica reprezentuje maksymalną możliwą wartość podobieństwa w przypadku, gdy wszystkie niejasności są rozstrzygnięte na korzyść podopieństwa. Szerokość przedziału może być interpretowana jako poziom niepewności procesu obliczania podobieństwa – im większa szerokość,

tym większa niepewność. Węższy przedział oznacza z kolei większą pewność obliczonego podobieństwa. Taka interpretacja jest w pełni zgodna z naszą interpretacją przypadku traktowania S1 jako dolnej granicy, a S4 jako górnej granicy. Dalsza analiza przypadków S2 i S3 może prowadzić do zmniejszenia górnej granicy i zwiększenia pewności w otrzymaną wartość podobieństwa. Jednakże analiza przypadków S2 i S3 jest obliczeniowo kosztowna.

Przedstawiona i opisana metoda obliczania podobieństwa była, i nadal jest, przedmiotem moich badań. Badania te dotyczą ulepszenia metody jak i jej zastosowania do rozwiązania rzeczywistych problemów.

W pracy [R1] metoda obliczania podobieństwa została poszerzona o możliwości porównania zasobów, których cechy reprezentowane są w różnych formatach danych [L35]. Metoda ta jest przystosowana do numerycznych jak i symbolicznych danych. Dodatkowo, dane te mogą być zlokalizowane w różnych miejscach na sieci. Metoda wykorzystuje elementy zbiorów rozmytych i agregacje lingwistyczną aby obliczyć podobieństwo biorąc pod uwagę wymagania i kryteria użytkowanika. Główną ideą proponowanej metody jest obliczanie podobieństwa każdej cechy porównywanych zasobów osobno i przekształcenie otrzymanych wartości w przestrzeń zbiorów rozmytych. Cechy mogą być wyrażone poprzez różne typy danych. Każdy typ danych jest związany z inną procedurą porównawczą. W pracy tej wykorzystujemy 2-tuple do reprezentacji wartości numerycznych [L36]. Pozwala to na porównanie i agregację lingwistycznych opisów cech. Zaproponowana metoda zastosowana została do wyszukiwania lekarstw na określoną dolegliwość. Eksperymenty przeprowadzono używając kilku zbiorów danych zawierających różne informacje dotyczące cech lekarstw. Znalezione lekarstwa spełniały dodatkowe wymagania odnośnie wartości niektórych cech. W przypadku, kiedy wartości te nie były podobne, metoda automatycznie wyszukiwała alternatywy z najlepszymi wartościami danej cechy. Dla porównania powtórzylismy proces szukania lekarstw używając języka SPARQL (ang. SPARQL Protocol and RDF Query Language) [L37]. Jednak ograniczenia języka SPARQL, na przyklad brak możliwości tworzenia zapytań niprecyzyjnych, uniemożliwiły nam powtórzenie wszystkich ekperymentów wykonanych używając proponowanej metody. Zastosowanie naszej metody pozwoliło na osiągnięcie lepszych rezultatów.

Proponowanej metodza obliczania podobieńtwa jest częścią innych prac badawczych. Praca [c42] poświecona jest problemowi asymilacji informacji. Zaproponowany algorytm oszacowywuje podobieństwo pomiędzy informacjami i integruje nową informację z informacją już znaną. W pracy [c55] rozważany jest proces budowy hierarchii kategorii. Jest to pierwszy krok w kierunku automatycznej konstrukcji definicji kategorii w oparciu o zebrane dane. Metoda przedstawiona w pracy jest sterowana danymi (ang. data-driven) i pozwala na określenie stopnia przynależności pojedyńczych danych do kategorii. Kontynuacją przedstawionych badań jest 'przetworzenie' skonstruowanych kategorii w ich definicje [c58]. Metody proponowane w tej pracy służą określeniu ważności własności kategorii oraz stopnia ich udziału w definicji. Odzwierciedla to rzeczywiste połączenia i zależności istniejące pomiędzy kategoriami. W ogólności, niektóre z tych połączeń są silne i znaczące, inne słabe i niezbyt istotne. Ważne badania dotyczące danych RDF koncentrują się również na automatycznym 'tłumaczniu' tekstu na zbiory trójek RDF [c52] oraz budowie algorytmów do konstrukcji czasowych zapytań nieprecyzyjnych [c46][c61].

V Literatura

(L1) P. De Meo, E. Ferrara, and G. Fiumara, Finding similar users in Facebook, *Social Networking and Community Behavior Modeling Qualitative and Quantitative Measurement*, IGI Global, 2011, pp. 304-323

(L2) I. Guy, M. Jacovi, A. Perer, I. Ronen, and E. Uziel, Same places, same things, same people? Mining

user similarity on social media, *Proc. ACM Conference on Computer Supported Cooperative Work*, Savannah, GA, 2010, pp. 41-50

(L3) A. Mathes, Folksonomies – Cooperative classification and communication through shared metadata, [Online]. Dostępne: http://www.adammathes.com/academic/computer-mediated-communication/ folksonomies.html (październik 27, 2018)

(L4) T. Hammond, T. Hannay, B. Lund, and J. Scott, Social bookmarking tools (I): A general review, D-Lib Mag. [Online]. 11 Dostępne: www.dlib.org/dlib/april05/hammond/04hammond.html (październik 27, 2018)

(L5) G. Smith, *Tagging: People-Powered Metadata for the Social Web*, New Riders, 2008

(L6) W.-T. Fu, T. Kannampallil, and R. Kang, A semantic imitation model of social tag choices, in *Proc. International Conference on Computational Science and Engineering – Volume 04*, 2009, pp. 66-73

(L7) L.A.Zadeh, Fuzzysets, *Information and Control*, 8,1965, pp. 338-353

(L8) R.R. Yager, Families of OWA operators, *Fuzzy Sets and Systems*, 59, 1993, pp. 125-148

(L9) R.R. Yager, Prioritized aggregation operators, *International Journal of Approximate Reasoning*, 48(1), 2008, pp. 263-274

(L10) B. Towle, and C. Quinn, Knowledge Based Recommender Systems Using Explicit User Models, Knowledge-Based Electronic Markets, *the AAAI Workshop*, Menlo Park, CA: AAAI Press, 2000, pp. 74-77

(L11) A. Sieg, B. Mobasher, R. Burke, Web Search Personalization with Ontological User Profiles, *Proc. 16th ACM Conference on Information and Knowledge Management*, Portugal, 2007, pp. 525-534

(L12) J. Trajkova, and S. Gauch, Improving ontology-based user profiles, *Proc. RIAO*, Vaucluse, France, 2004, pp. 380-389

(L13) G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, *IEEE Trans Knowledge and Data Engineering*, 17(6), 2005, pp. 734-749

(L14) R. Burke, Hybrid recommender systems: survey and experiments, *User Model User Adapt Interact*, 2002, 12(4), pp. 331-370

(L15) J.B. Schafer, J.A. Konstan, J. Riedl, Recommender systems in e-commerce, *ACM Conference on Electronic Commerce*, Denver, CO, 1999, pp. 158-166

(L16) K. Lakiotaki, P. Delias, V. Sakkalis, N.F. Matsatsinis, User profiling based on multi-criteria analysis: the role of utility functions, *International Journal on Operational Research*, 9, 2009, pp. 3-16

(L17) J. Freyne, S. Berkovsky, E. M. Daly, and W. Geyer, Social networking feeds: Recommending items of interest, *ACM Conference on Recommender Systems*, Barcelona, Spain, 2010

(L18) G. Adomavicius, Y.O. Kwon YO, New recommendation techniques for multicriteria rating systems, *IEEE Intelligent Systems*, 22(3), 2007, pp. 48-55

(L19) L.A. Zadeh, Fuzzy logic = computing with words, *IEEE Transactions on Fuzzy Systems*, 4, 1996, pp. 103-111

(L20) F. Sebastiani, Machine learning in automated text categorization, *ACM Computing Surveys*, 34(1), 2002, pp.1-47

(L21) V. Vidulin, M. Lustrek, M. Gams, Training a genre classifier for automatic classification of web pages, *Journal of Computing and Information Technology*, 15(4), 2007, pp. 305-311

(L22) O. Dridi, M.B. Ahmed, Building an ontology-based framework for semantic information retrieval: application to breast cancer, *Proc. International Conference on Information and Communication Technologies: from Theory to Applications*, Damascus, Syria, 2, 2008, pp. 1-6

(L23) P. Castells, M. Fernandez and D. Vallet, An Adaptation of the Vector- Space Model for Ontology-Based Information Retrieval, *IEEE Transactions on Knowledge and Data Engineering*, 19(2), 2007, pp. 261-272

(L24) S. L. Tomassen, Searching with Document Space Adapted Ontologies, *Emerging Technologies and Information Systems for the Knowledge Society*, 5288, 2008, pp. 513-522

(L25) R.R.Yager, A Hierarchical Document Retrieval Language, *Information Retrieval*, 3, 2000, pp. 357-377

(L26) O. Lassila, R. Swick, Resource description framework (RDF) model and syntax specification, *World Wide Web Consortium Technical Reports and Publications*, Available: https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/ (Accessed October 27, 2018)

(L27) N. Shadbolt, W. Hall, T. Berners-Lee, The semantic web 1322 revisited. *IEEE Intelligent Systems*, 21, 2006, pp. 96-101

(L28) P. Resnik, Semantic Similarity in a Taxonomy: An Information-Based, *Journal of Artificial Intelligence Research*, 11, 1999, pp. 95-130

(L29) M. A. Rodriguez, and M. J. Egenhofer, Determining Semantic Similarity among Entity Classes from Different Ontologies, *IEEE Transactions on Knowledge and Data Engineering*, 15, 2003, pp. 442-456

(L30) A. Tversky, Features of similarity, *Psychol Review*, 84, 1977, pp. 327-352

(L31) R.M. Nosofsky, Stimulus bias, asymmetric similarity, and classification, *Cognitive Psychology*, 23, 1991, pp. 94-140

(L32) C. Bizer, T. Heath, T. Berners-Lee, Linked data-the story so far, *International Journal on Semantic Web and Information Systems* 4, 2009, pp. 1-22

(L33) D. Dubois, H. Prade, *Possibility theory: an approach to computerized processing of uncertainty*, Plenum press, New 1251 York, 1988

(L34) D. Dubois, H. Prade, Possibility theory and its applications: a retrospective and prospective view, *FUZZ-IEEE International Conference on Fuzzy Systems*, St. Louis, MO, 2003.

(L35) L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility, *Fuzzy Sets and Systems*, 1, 1978, pp. 3-28

(L36) F. Herrera, L. Martinez, A 2-tuple fuzzy linguistic representation model for computing with words, *IEEE Transactions on Fuzzy Systems*, 8, 2000, pp. 746-752

(L37) B., Quilitz, U. Leser, Querying distributed RDF data sources with SPARQL, *5th European Semantic Web Conference - ESWC*, 2008, pp. 524-538