Autoreferat

1 Imię i Nazwisko: Marcin Sydow

2 Posiadane dyplomy, stopnie naukowe:

- magister matematyki,
  Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski, 1997
- doktor nauk technicznych w zakresie informatyki,
  Instytut Podstaw Informatyki, Polska Akademia Nauk, 2004:
  "Link Analysis of the Web Graph. Measurements, Models and Algorithms for Web Information Retrieval"

3 Informacje o dotychczasowym zatrudnieniu w jednostkach naukowych:

- Instytut Podstaw Informatyki, Polska Akademia Nauk, adiunkt, od 2010
- Polsko-Japońska Wyższa Szkoła Technik Komputerowych, adiunkt, od 2004
- Polsko-Japońska Wyższa Szkoła Technik Komputerowych, asystent, 1999-2004
- Politechnika Warszawska, Wydział Fizyki Teoretycznej i Matematyki Stosowanej, asystent, 1997-1998

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:

"Dywersyfikacja informacji w wybranych zadaniach selekcji, wyszukiwania i prezentacji informacji w semantycznych grafach wiedzy i wyszukiwarkach internetowych"

4.2 Cykl 6 publikacji stanowiących osiągnięcie naukowe:

1. **[SPS13] M.Sydow**, M.Pikuła, R.Schenkel **"The Notion of Diversity in Graphical Entity Summarisation on Semantic Knowledge Graphs"** Journal of Intelligent Information Systems, Volume 41, Issue 2, pp. 109-149, ISSN 0925-9902, DOC: 10.1007/s10844-013-0239-6, Springer, 2013.
(na liście **JCR** i w bazie **Web of Science**, Impact Factor: **0.632**)

   Wkład własny: **70%** (M.Pikuła: 20%, R.Schenkel: 10%)

   20 cytowań (Google Scholar), **2 cytowania** (Web of Science)

2. **[KKRS13]** W.Kosiński, T.Kuśmierczyk, P.Rembelski, **M.Sydow "Application of Ant-Colony Optimisation to Compute Diversified Entity Summarisation on Semantic Knowledge Graphs"**, Annals of Computer Science and Information Systems, Volume 1, pp. 69-76, ISSN 2300-5963, (also: Proc. of International IEEE AAIA 2013/FedCSIS Conference, ISBN 978-1-4673-4471-5), PTI/IEEE, 2013. (obecna w bazie **Web of Science**)

   Wkład własny: **40%** (W.Kosiński,T.Kuśmierczyk po 25%, P.Rembelski 10%)

3. **[Syd14a] M.Sydow "Approximation Guarantees for Max Sum and Max Min Facility Dispersion with Parameterised Triangle Inequality and Applications in Result Diversification"** (extended journal version) Mathematica Applicanda vol. 42(2), pp. 241–257, Print ISSN: 1730-2668, On-line ISSN: 2299-4009, DOI: 10.14708/ma.v42i0.547, PTM, 2014.

   Wkład własny: **100%**

4. **[MSS14]** S.Metzger, R.Schenkel, **M.Sydow "Aspect-based Similar Entity Search in Semantic Knowledge Graphs with Diversity-awareness and Relaxation"** Proceedings of the IEEE/WIC/ACM WI-IAT 2014 (Intern. Joint Conference on Web Intelligence and Intell. Agent Techn.), pp. 60-69, ISBN 978-1-4799-4143-8, DOI 10.1109/WI-IAT.2014.17, IEEE Computer Society, 2014.

   Wkład własny: **25%** (S.Metzger: 50%, R.Schenkel: 25%)

   1 cytowanie (Google Scholar)

5. **[SBCD09] M.Sydow**, F.Bonchi, C.Castillo, D.Donato **"Optimising Topical Query Decomposition"** Proc. of WSCD09 (collocated with ACM WSDM 2009 Conference) pp. 43-47, ISBN 978-1-60558-434-8,
DOI: 10.1145/1507509.1507516, ACM, 2009.

   Wkład własny: **65%** (F.Bonchi 15%, C.Castillo 10%, D.Donato 10%)

   5 cytowań (Google Scholar)

6. **[SCW12] M.Sydow**, K.Ciesielski, J.Wajda **"Introducing Diversity to Log-based Query Suggestions to Deal with Underspecified User Queries"**, LNCS Vol. 7053, pp. 251-264 (Proc. of Intern. SIIS 2011 Conference, Revised Selected Papers), ISBN 978-3-642-25260-0, Springer, 2012. (**Web of Science**) Wkład własny: **55%** (K.Ciesielski: 20%, Wajda: 25%)

   6 cytowań (Google Scholar), **2 cytowania** (Web of Science)

4.3 Ogólne omówienie celu naukowego ww. prac, wyników oraz ich zastosowań.

Przedmiotem osiągnięcia jest stworzenie i eksperymentalna ewaluacja nowych technik tzw. *dywersyfikacji* informacji, co ma zastosowanie w interaktywnych systemach informacyjnych, w których:

- system ma dokonać selekcji i prezentacji użytkownikowi *niewielkiego* zbioru reprezentatywnych informacji, tak aby najlepiej spełnić jego potrzebę informacyjną;
- system musi dokonać selekcji spośród dużej liczby obiektów potencjalnie spełniających potrzebę użytkownika, przy czym istnieje wiele możliwych interpretacji potrzeby informacyjnej użytkownika.

Istotne jest to, że liczba obiektów, które należy wybrać i zaprezentować jest niewielka w porównaniu z liczbą obiektów potencjalnie spełniających potrzebę użytkownika. Przykładem takiego systemu jest wyszukiwarka WWW, gdzie użytkownik formułuje zwykle krótkie i wieloznaczne zapytanie oczekując krótkiej listy (np. 10) dokumentów WWW, które w jak największej części *trafnie* spełnią jego potrzebę informacyjną wyrażoną przez zapytanie, przy czym liczba dostępnych dokumentów jest ogromna (np. rzędu 10^9) a zapytanie może być interpretowane na wiele sposobów. Innym przykładem jest system automatycznie *podsumowujący* tekst, gdzie zadaniem jest wygenerowanie krótkiego podsumowania składającego się z niewielkiej w porównaniu z wyjściowym tekstem liczby zdań, które reprezentatywnie podsumują jak najwięcej jego aspektów.

W zadaniach tego typu znajdują zastosowanie techniki polegające na *celowej i kontrolowanej dywersyfikacji* zwracanych wyników, tak aby niewielki zwracany użytkownikowi zbiór reprezentował jak najwięcej możliwych aspektów lub interpretacji potrzeby informacyjnej użytkownika, aby w efekcie w tym niewielkim zbiorze zawrzeć obiekty, które *faktycznie* okażą się trafne dla użytkownika.

Nowatorskość osiągnięcia polega m.in. na *rozwinięciu i adaptacji* istniejących technik dywersyfikacji do nowych zadań, w których techniki takie nie były do tej pory stosowane, a także na zaproponowaniu nowych pomysłów. W szczególności, w pracach innych badaczy dywersyfikacji podlegały do tej pory głównie obiekty *tekstowe* (dokumenty, zdania) natomiast zaproponowane przez autora rozwiązania dotyczą zupełnie innych i czasem nowych obiektów takich jak np. elementy *semantycznych grafów wiedzy* lub *podpowiedzi zapytań* w wyszukiwarkach WWW. W związku z tym zaprojektowano też zupełnie nowe elementy technik, adekwatne do nowych danych i zadań. Stworzone techniki dywersyfikacji dotyczą następujących nowych zastosowań:

- *zdywersyfikowane podsumowywanie encji w semantycznych grafach wiedzy* [SPS13, KKRS13, SPS12, Syd14c, SPS11, SPS10b, Syd11, SPS10a, SPSS10]
- *wybór encji podobnych do zadanych* (w semantycznych grafach wiedzy) [SCSS15, MSS13, MSS14]
- automatyczne generowanie zdywersyfikowanych *podpowiedzi zapytań* w wyszukiwarkach internetowych [SMN+15, SCW12, SBCD09, MNSS13].

Spośród wybranych 6 prac stanowiących osiągnięcie, 3 dotyczą semantycznych grafów wiedzy. Przy czym, spośród tych 3, 2 prace [SPS13, KKRS13] dotyczą zdywersyfikowanych *podsumowań encji*, a 1 praca [MSS14] zdywersyfikowanego wyboru (sugerowania użytkownikowi) *zbioru encji podobnych do zadanych* encji wejściowych. Następnie, 2 prace [SCW12, SBCD09] dotyczą *generowania zdywersyfikowanych podpowiedzi zapytań* w wyszukiwarkach internetowych, natomiast 1 praca [Syd14a] zawiera pewne pomocnicze *wyniki teoretyczne*, mające potencjalne zastosowania we wszystkich technikach dywersyfikacji, które mogą być wywiedzione z problemu *dyspersji* (ang. *facility dispersion*), w szczególnosci w rozważanych w pozostałych pracach cyklu.

Wymienione 6 publikacji stanowiących osiągnięcie naukowe stanowi jedynie wybór reprezentujący znacznie liczniejszy zbiór około 20 prac autora dotyczących tematyki dywersyfikacji. Lista pozostałych prac autora z lat 2008-2015 pozwiązanych z tematyką dywersyfikacji wymieniona jest w sekcji 5. Prace o tej tematyce stanowią około 30% całego dorobku publikacyjnego (ok. 65 prac) autora. Autor niezależnie od tego zajmował się licznymi innymi tematami badawczymi, co opisane jest w sekcji 5.

4.4 Krótkie omówienie poszczególnych prac stanowiących osiągnięcie

4.4.1 Zdywersyfikowane podsumowania encji w semantycznych grafach wiedzy (praca [SPS13])

Podstawowy wątek osiągnięcia reprezentuje publikacja [SPS13], gdzie opisany jest oryginalnie zaproponowany przez autora problem *grafowego podsumowania encji* w semantycznych grafach wiedzy (ang. *Graphical Entity Summarisation – GES*) i jednocześnie zaproponowane i badane jest podejście oparte na *dywersyfikacji* do tego problemu – DIVERSUM (ang. *diversified entity summarisation*).

Termin *"encja"* (ang. *entity*) jest podstawowym terminem z dziedziny baz danych, stanowiącym reprezentację w bazie danych abstrakcyjnego lub rzeczywistego obiektu z modelowanej dziedziny (np. osoba, kraj, partia polityczna, etc.).

Natomiast *podsumowanie* encji, poprzez analogię do np. podsumowywania tekstu, stanowi proces selekcji (spośród wielu dostępnych w bazie) niewielkiego zbioru reprezentatywnych informacji dotyczących danej encji i prezentacji ich w zwartej formie użytkownikowi. W przypadku semantycznych grafów wiedzy, informacje dotyczące danej encji nazywane są też *faktami*. Fakt w takiej bazie reprezentowany jest przez parę węzłów (reprezentujących encje) połączoną krawędzią (reprezentującą jakąś relację między encjami, np. ("Chopin", *"pochodzi z"*, "Polski").

Zazwyczaj całkowita liczba faktów dostępnych w bazie dotyczących danej encji (np. słynnego polityka, państwa, etc.) jest zbyt duża do łatwego przyswojenia dla użytkownika, więc podsumowanie encji (wyselekcjonowany zbiór faktów) stanowi użyteczne narzędzie we wszelkich systemach analitycznych, czy służących do przeglądania lub wyszukiwania informacji w semantycznych bazach danych.

Powyżej opisane zadanie spełnia dwa warunki opisane na początku sekcji 4.3 stanowi więc naturalne pole zastowań dla technik dywersyfikacji.

Obliczone przez system podsumowanie zwracane jest w formie *grafu* i przedstawiane użytkownikowi w formie graficznej, co stanowi nowość, gdyż w istnieją-

cych wcześniej systemach dotyczących semantycznych grafów wiedzy, wynik prezentowany był w formie elementów tekstu (np. wyszukiwarka NAGA [KSI+08] dla systemu YAGO [yag]). Wszystkie wymienione wyżej cechy badanego problemu sprawiają, że praca [SPS13] jest oryginalną syntezą i rozwinięciem kilku innych niezależnie istniejących pomysłów i stanowi pionierską pracę dotyczącą tak ujętego problemu.

Zawartość pracy [SPS13]: We wstępnych częściach pracy [SPS13] omówiona jest motywacja, podany przykładowy konkretny scenariusz zastosowania praktycznego proponowanego rozwiązania i wyjaśnienie modelu semantycznego grafu wiedzy. Dokonany jest też przegląd literatury dotyczącej dziedzin, których syntezę stanowi ta praca: podsumowań tekstów, i istniejących technik dywersyfikacji informacji w innych dziedzinach. Omówiony jest też kontekst zagadnienia wyszukiwania informacji (ang. IR - *information retrieval*) stanowiący inspirację dla oryginalnych idei autora przy rozwiązywaniu omawianego nowego problemu.

Następnie opisane są 2 zaproponowane przez autora algorytmy pozwalające efektywnie rozwiązać proponowany problem. Pierwszy algorytm (nazwany PRECIS) pozwala efektywnie obliczać zadane podsumowanie encji uwzględniając 2 kryteria doboru faktów (krawędzi grafu semantycznego): *trafność* (ang. *relevance*) i *ważność* (ang. *importance*). Algorytm PRECIS nie uwzględnia kryterium dywersyfikacji i jest użyty w dalszej części pracy jako algorytm referencyjny do porównania wyników z algorytmem uwzględniającym dywersyfikację. Algorytm jest autorskim rozwinięciem i adaptacją klasycznego algorytmu Dijkstry znajdowania najkrótszych ścieżek z jednym źródłem w grafach [Dij59] na przypadek semantycznych (multi-)grafów wiedzy i z odpowiednim rozbudowaniem pojęcia odległości. W tym zachłannym algorytmie elementy podsumowania dodawane są sukcesywnie aż do uzyskania zadanego rozmiaru w kolejności od "najbliższych" do "najdalszych" w sensie odpowiednio zdefiniowanej odległości od podsumowywanego węzła. Następnie opisany jest drugi zaproponowany przez autora efektywny algorytm (nazywany w tej pracy DIVERSUM) dla rozważanego problemu. Algorytm ten uwzględnia 4 kryteria doboru faktów do podsumowania: *trafność*, *dywersyfikacja*, *popularność* i *ważność*.

Następnie przedstawiono wyniki *obiektywnej* eksperymentalnej ewaluacji na danych z semantycznego grafu wiedzy YAGO [yag] przy użyciu zaproponowanej miary jakości oraz innego eksperymentu ewaluacyjnego porównującego wyniki obu algorytmów przy użyciu informacji dostępnej w serwisie Wikipedia [wik]. Do ewaluacji porównawczej użyto algorytmu PRECIS jako rozwiązania referencyjnego.

Oprócz tego przedstawiono wyniki kolejnej serii eksperymentów *subiektywnej* ewaluacji z użyciem opinii ekspertów oraz z udziałem użytkowników, w których zastosowano techniki tzw. *crowdsourcingu*. Technika ta polega na masowym zlecaniu wykonywania drobnych zadań niezidentyfikowanej szerokiej grupie ludzi za pośrednictwem specjalnego portalu WWW, np. *mechanical turk* [mtu]. Dla zwiększenia wiarygodności wyników eksperymentów subiektywnych, ukryto przed użytkownikami oceniającymi algorytmy fakt stosowania podejścia opartego na dywersyfikacji w jednym z algorytmów a wyniki były przedstawiane w losowej kolejności. Dodatkowo, zastosowano opcjonalne pole tekstowe, w którym użytkownicy mogli wyrazić w formie otwartych komentarzy opinie o ocenianych wynikach. Użytkown-

icy samodzielnie podkreślali w owych komentarzach, że preferują wyniki bardziej *zróżnicowane*.

We wszystkich wymienionych eksperymentach *wykazano przewagę podejścia uwzględniającego dywersyfikację* wyników. Potwierdza to główną tezę pracy, że dywersyfikacja wyników w badanym problemie jest podejściem pożądanym, poprawia jakość wyników i zwiększa satysfakcję użytkowników z działania systemu.

Jednocześnie eksperymenty w pracy [SPS13] ujawniły jako dodatkowy wynik, że stopień dywersyfikacji wyników zwracanych przez algorytm DIVERSUM jest potencjalnie *zbyt wysoki*, mimo ogólnie pozytywnego efektu zastosowania dywersyfikacji do problemu GES. Dokładniej, poprzez dychotomiczne podejście, tj ścisłe unikanie redundancji w algorytmie DIVERSUM, niemożliwe jest subtelne odzwierciedlenie rozkładu relacji będących udziałem podsumowywanej encji. W konkluzji pracy [SPS13] zauważa się więc, że dalsze ulepszenie algorytmu DIVERSUM polegałoby na wprowadzeniu parametru, dzięki któremu możliwe byłoby kontrolowanie *stopnia dywersyfikacji*. W takim ujęciu, poprzednie podejścia – brak dywersyfikacji (algorytm PRECIS) i maksymalna dywersyfikacja (algorytm DIVERSUM, unikający jakichkolwiek powtórzeń) – postrzegane być mogą jako skrajne podejścia i mimo przewagi drugiego z nich, pożądane byłoby rozwiązanie pośrednie, najlepiej z możliwością stopniowania.

Praca nad publikacją [SPS13] rozwijana była w intensywnej współpracy z Max Planck Institut fuer Informatik, w Saarbruecken, w Niemczech w ramach zespołu, który rozwija m.in. semantyczny graf wiedzy YAGO [yag]. Autor odbył też w jej ramach kilka staży/wizyt naukowych w tym ośrodku w latach 2009, 2010 i 2013. Na finansowanie jednego z nich autor uzyskał stypendium DAAD w roku 2010.

Wyniki wcześniejszych etapów pracy nad publikacją [SPS13] prezentowane były na licznych konferencjach międzynarodowych takich jak ACM CIKM 2009, IEEE ICDE 2010, IMCSIT 2010, ECIR 2011, ISMIS 2012.

Praca [SPS13] poprzez postawienie i zarazem rozwiązanie nowego problemu w krótkim czasie od ukazania się doczekała się cytowań (ok. 25 cytowań) na konferencjach i w czasopismach specjalistycznych dotyczących tej problematyki.

Wniosek dotyczący tematyki badawczej związanej z pracą [MSS14] wygrał konkurs NCN na finansowanie grantem N N516 481940, którego autor był kierownikiem.

4.4.2 DIVERSUM jako problem optymalizacyjny (praca [KKRS13])

Ze względu na obserwację wymienioną pod koniec sekcji 4.4.1, w pracy [KKRS13], która stanowi część opisywanego osiągnięcia i jest naturalną kontynuacją pracy [SPS13], zaproponowano ulepszony algorytm DIVERSUM, w którym stopień dywersyfikacji może być kontrolowany za pomocą parametru $\sigma \in R^+$ w ten sposób, że im większa jego wartość tym bardziej zdywersyfikowane są wyniki (dla $\sigma = 0$ algorytm nie uwzględnia dywersyfikacji i zachowuje się podobnie do algorytmu PRECIS).

W tym celu zaadaptowano do problemu GES techniki dywersyfikacji wyników stosowane nieco wcześniej przez innych autorów [GS09], ale w dziedzinie wyszukiwania dokumentów www, a oparte na badanym jeszcze wcześniej w dziedzinie badań operacyjnych problemie *dyspersji* (ang. *Facility Dispersion – FD*) [CH96].

W problemie FD chodzi o rozmieszczenie przestrzenne $k \in N^+$ obiektów poprzez wybranie spośród wielu potencjalnych lokalizacji niewielkiego zbioru lokalizacji tak, aby były one możliwie "rozrzucone" w sensie odległości. Podane są odległości parami wszystkich potencjalnych lokalizacji. Stopień rozrzucenia przestrzennego wybranych lokalizacji mierzony jest odpowiednią funkcją celu a zadanie sformułowane jest jako optymalizacja (maksymalizacja) tej funkcji. W pracy [GS09] autorzy zaproponowali postawienie problemu dywersyfikacji wyników wyszukiwania w sieci WWW jako problemu optymalizacyjnego będącego niejako rozszerzeniem dwóch wariantów problemu Facility Dispersion.

W pracy [KKRS13] autor zaproponował z kolei adaptację techniki dywersyfikacji wyników wyszukiwania przedstawioną przez innych badaczy w [GS09] do problemu GES, czyli do zaprojektowania ulepszonej wersji algorytmu DIVERSUM.

W pracy [KKRS13] zdefiniowano problem zdywersyfikowanego podsumowywania encji w semantycznych grafach wiedzy rozszerzając specyfikację problemu GES o założenie, że dostępna jest funkcja "ważności" imp (ang. *importance*) danego faktu w kontekście podsumowywanej encji q oraz funkcja $diss$ (ang. *dissimilarity*) "niepodobieństwa" par faktów podsumowujących encję q. Przy takich założeniach zdefiniowano na nowo ten problem jako optymalizację następującej funkcji celu:

$$obj_q(S) = (k-1)\sum_{d \in S} imp_q(d) + 2\sigma \sum_{d_1 \neq d_2 \in S} diss_q(d_1, d_2), \qquad (1)$$

która wyraża parametryzowaną kombinację łącznej wartości "ważności" oraz "zdywersyfikowania" danego zbioru faktów S.

Celem jest znalezienie zbioru S zawierającego k faktów, który maksymalizuje wartość funkcji celu $obj_q(S)$. Pierwszy składnik we wzorze (1) odpowiada za sumę ważności wybranych faktów, natomiast drugi za ich wzajemne sumaryczne "niepodobieństwo parami". Intuicja jest taka, że maksymalizując drugi człon eliminujemy redundancję czyli zbyt duże wzajemne podobieństwo wybranych elementów, co równoważne jest ich "zdywersyfikowaniu". Współczynniki $k-1$ oraz 2 mają za zadanie odzwierciedlić fakt, że istnieje dokładnie $k(k-1)/2$ par różnych elementów ze zbioru k-elementowego.

W ten sposób zdefiniowano nowy w porównaniu do [SPS13] sposób otrzymywania zdywersyfikowanych podsumowań encji w semantycznych grafach wiedzy (jako optymalizację w/w funkcji), w którym możliwe jest kontrolowanie stopnia dywersyfikacji za pomocą parametru σ tak jak planowano w konkluzji pracy [SPS13].

Optymalizacja tak zdefiniowanej funkcji celu jest równoważna optymalizacji w zadaniu Max Sum Facility Dispersion (co pokazano m.in. w [GS09]), a więc w sensie złożoności obliczeniowej jest zadaniem NP-trudnym. Wobec tego w pracy [KKRS13] autor niniejszego autoreferatu w celu efektywnego rozwiązania tak postawionego trudnego zadania obliczeniowego zaproponował jako narzędzie obliczeniowe heurystykę *optymalizacji mrówkowej* ACO (ang. *Ant Colony Optimisation*) do znajdowania rozwiązania przybliżonego.[1]

Zagadnieniem interesującym naukowo samym w sobie jest sposób obliczania funkcji imp (ważności pojedynczego faktu w kontekście podsumowywanej encji q

[1] Zagadnienia teoretyczne i implementacyjne samej heurystyki ACO nie stanowią części niniejszego osiągnięcia naukowego i zostały wykonane przez współautorów publikacji [KKRS13].

w problemie GES) oraz $diss$ (niepodobieństwa parami dwóch faktów w kontekście podsumowywanej encji).[2] Zagadnienie obliczania tych funkcji zasługuje na oddzielne badania i zakłada się, że funkcje imp oraz $diss$ są dane. Jednak na potrzeby pracy [KKRS13] autor zaproponował tutaj dla funkcji imp algorytm błądzenia losowego w grafie wiedzy z powrotami w oparciu o intuicję, że fakty "ważne" dla danej encji q to te, które będą często "odwiedzane" w procesie błądzenia rozpoczętego w wierzchołku q i z powrotami do tego wierzchołka, z uwagi na bliskość topologiczną. Funkcję $diss$ z kolei obliczano na podstawie statystyk współwystępowania w grafie faktów (krawędzi) o danych etykietach w incydencji z różnymi encjami (zgodnie z intuicją, że im rzadziej krawędzie o danych etykietach są współincydentne z daną encją tym bardziej są "niepodobne" do siebie).

Następnie opisane wyżej podejście zaimplementowano i dokonano ewaluacji empirycznej na zbiorze danych YAGO [yag]. W tym celu zdefiniowano kilka obiektywnych miar będących adaptacją miar precyzji (ang. *precision*) i pełności (ang. *recall*) do rozważanego problemu. Porównano wyniki tego nowego podejścia do wyników algorytmów PRECIS i DIVERSUM oraz do zbioru referencyjnego stworzonego na podstawie Wikipedii.

Wg wszystkich wykonanych eksperymentów przedstawionych w [KKRS13] nowo zaproponowane podejście do dywersyfikacji podsumowań encji, polegające na maksymalizacji funkcji opisanej równaniem (1), dla pewnej wartości parametru σ, było w stanie osiągnąć najlepsze wyniki spośród wszystkich testowanych algorytmów, w tym PRECIS i DIVERSUM, omówionych wcześniej. Co ważne, wyniki algorytmu DIVERSUM mogły być polepszone przez nowy algorytm tylko dla odpowiednio wysokiej wartości parametru σ, czyli przy odpowiednio wysokim stopniu dywersyfikacji wyników.

Eksperymenty te jeszcze raz potwierdziły, że wprowadzenie nowego mechanizmu dywersyfikacji umożliwiło dalsze (w stosunku do algorytmu DIVERSUM z pracy [SPS13]) polepszenie jakości wyników wg wszelkich przyjętych obiektywnych miar ewaluacyjnych w problemie podsumowania encji w semantycznych grafach wiedzy.

Wyniki przedstawione w pracy [KKRS13] prezentowane były na międzynarodowej konferencji IEEE AAIA/FedCSIS 2013.

Wniosek dotyczący tematyki badawczej związanej z pracą [MSS14] wygrał konkurs NCN na finansowanie grantem N N516 481940, którego autor był kierownikiem.

4.4.3 Wyniki dotyczące efektywnej aproksymacji rozwiązań problemów dywersyfikacji sformułowanych na bazie problemu Facility Dispersion (praca [Syd14a])

Kontynuacją pracy [KKRS13], w której zdefiniowano zdywersyfikowany problem GES jako optymalizację funkcji opisanej równaniem (1), jest publikacja [Syd14a]. W pracy tej bada się pewne algorytmiczne aspekty problemu efektywnego znajdowania przybliżonych rozwiązań dla zadania Max Sum i Max Min Facility Dispersion. Ponieważ, jak przypomniano w sekcji 4.4.2, problem obliczania zdywersyfikowanych podsumowań encji w formie rozważanej w [KKRS13] jest algorytmicznie równoważny problemowi Max Sum Facility Dispersion, wyniki pracy

[2] Przypomnieć przy tym należy przyjęte w tej pracy ważne założenie o korzystaniu wyłącznie z danych zawartych w danym grafie wiedzy, bez możliwości korzystania z danych zewnętrznych.

[Syd14a] mają potencjalne zastosowania praktyczne w problemie będącym przedmiotem opisywanego osiągnięcia. Co więcej, wyniki te mają potencjalne zastosowania w wielu innych zagadnieniach związanych z dywersyfikacją w sformułowaniach wywiedzionych z problemów Max Sum i Max Min Facility Dispersion, w szczególności w problemie *dywersyfikacji wyników wyszukiwania www* (ang. *result diversification*) [GS09].

Dokładniej, w pracy [Syd14a] dowodzi się, że NP-trudne problemy Max Sum i Max Min Facility Dispersion można w czasie wielomianowym aproksymować[3] ze współczynnikiem $\frac{2}{\alpha}$, gdy funkcja odległości $d(\cdot,\cdot)$ spełnia następującą *parametryzowaną nierówność trójkąta*:

$$d(x,y) + d(y,z) \geq \alpha d(x,z),$$

dla każdej trójki elementów x, y, z rozważanego zbioru uniwersalnego. Wyniki te uogólniają znane wcześniej wyniki dotyczące przypadku zwykłej nierówności trójkąta, tj $\alpha = 1$, na dowolną wartość $\alpha \in [0, 2]$ czyli w zakresie od tzw "metryki dyskretnej" do "semi-metryki".

Co ważne, dodatkową wartość wyników pracy [Syd14a] stanowi zawarte w niej częściowe rozwiązanie problemu naukowego postawionego niedawno przez innego naukowca w pracy prezentowanej na międzynarodowej konfenecji PODS 2012 [BLY12]. Problem postawiony na końcu tej pracy sformułowany został następująco: *"czy możliwe jest odniesienie współczynnika aproksymacji do parametru uogólnionej nierówności trójkąta, którą spełnia funkcja odległości w tym problemie?"* (tłum. z ang. autora). Omawiana praca [Syd14a] w sposób jasny daje odpowiedź na to pytanie w szczególnym przypadku (autor [BLY12] rozważał ogólniejszy przypadek).

Wyniki teoretyczne uzyskane w [Syd14a] mają zastosowania nie tylko w problemach zdywersyfikowanych podsumowań encji, ale także w dużo szerszym zbiorze zagadnień dywersyfikacji, które potencjalnie wyspecyfikować można na bazie problemu dyspersji (Facility Dispersion). Do takich potencjalnych zastosowań zaliczyć można np. systemy rekomendacyjne, wyszukiwarki WWW czy systemy do podsumowywania informacji.

Wyniki wczesnych etapów pracy przedstawionej w publikacji [Syd14a] prezentowane były na konferencji międzynarodowej ISMIS 2014, oraz krajowej KZM 2014.

Wniosek dotyczący tematyki badawczej związanej z pracą [MSS14] wygrał konkurs OPUS 5 NCN na finansowanie grantem 2012/07/B/ST6/01239, którego autor jest kierownikiem.

4.4.4 Zagadnienie dywersyfikacji w problemie znajdowania encji podobnych w semantycznych grafach wiedzy (praca [MSS14])

Kolejny wątek związany z zastosowaniami dywersyfikacji wyników stanowi przedmiot publikacji [MSS14]. W pracy tej rozważa się problem zwracania zdywersyfikowanego zbioru encji podobnych do zadanych (przez użytkownika) w oparciu o semantyczny graf wiedzy. Jest to więc wątek pokrewny do tego opisywanego w

[3] tzn. jest matematyczna gwarancja, że wartość funkcji celu znalezionego przybliżonego rozwiązania będzie się różnić od optymalnego conajwyżej o stały współczynnik multiplikatywny

sekcjach 4.4.1 i 4.4.2, gdyż dotyczy podobnych obiektów (encji) i tej samej formy danych (semantyczny graf wiedzy). W pracy tej wprowadza się nowy, 3-poziomowy model reprezentacji encji oparty na strukturalnych własnościach grafu wiedzy i bazujący na nim nowy algorytm znajdowania encji podobnych zapewniający dywersyfikację wyników. Proponuje się też algorytmy rangowania wyników i prezentuje eksperymentalne wyniki porównujące zaproponowane podejście do szeregu istniejących algorytmów. Eksperymenty dokonane są na dostępnych publicznie zbiorach danych, co gwarantuje ich powtarzalność.

Główny wkład autora w pracę [MSS14] obejmuje wyspecyfikowanie problemu, zaproponowanie w/w nowego, 3-poziomowego modelu charakteryzacji encji i opartego na nim nowego algorytmu doboru encji podobnych w oparciu o tzw. *aspekty maksymalne*. Pomysł ten zapewnia, że znalezione encje są jednocześnie *maksymalnie podobne* do zadanych w sensie zaproponowanego modelu reprezentacji strukturalnej, a z drugiej strony są *zdywersyfikowane* w tym sensie, że zbiory encji odpowiadających dowolnym dwóm różnym aspektom maksymalnym są rozłączne, a więc automatycznie unika się redundancji. Wykazane to zostało w postaci zamieszczonego w pracy twierdzenia wraz z dowodem.

Zaprezentowane w [MSS14] liczne wyniki eksperymentalne wykonane na kilku publicznie dostępnych zbiorach danych pokazują wyższość proponowanego rozwiązania nad istniejącymi rozwiązaniami konkurencyjnymi innych autorów.

Wkład autora w pracę [MSS14] jest kluczowy, gdyż definiuje koncepcje i główny mechanizm znajdowania zdywersyfikowanego zbioru encji podobnych do zadanych, który zapewnia bardzo dobre działanie zaimplementowanego algorytmu, mimo faktu, że ilościowo wkład jednego ze współautorów jest większy, gdyż obejmuje wszelkie prace implementacyjne i dotyczące rangowania encji, wykonania, prezentacji i analizy eksperymentów.

Prace przedstawione w publikacji prowadzono we współpracy z Max-Planck Institut fuer Informatik, w Saarbruecken, w Niemczech i w związku z nimi odbyto staż naukowo-badawczy w tym ośrodku w roku 2013.

Wyniki pracy [MSS14] i jej wcześniejszych etapów prezentowane były na międzynarodowych konferencjach ACM CIKM 2013 oraz IEEE/WIC/ACM WI-IAT 2014.

Wniosek dotyczący tematyki badawczej związanej z pracą [MSS14] wygrał konkurs OPUS 5 NCN na finansowanie grantem 2012/07/B/ST6/01239, którego autor był kierownikiem.

4.4.5 Generowanie zdywersyfikowanego zbioru podpowiedzi zapytań poprzez tematyczną dekompozycję zapytania (praca [SBCD09])

Ostatni reprezentowany w osiągnięciu wątek dotyczy specjalistycznego problemu automatycznego generowania zdywersyfikowanego zbioru *podpowiedzi* po podaniu przez użytkownika wyszukiwarki internetowej *niejednoznacznego zapytania*. Celem takiego zadania jest ułatwienie użytkownikowi szybszego dotarcia do satysfakcjonujących go wyników wyszukiwania poprzez wybór takiej podpowiedzi, która najlepiej *ujednoznaczni* (uściśli) aspekt jego wieloznacznego zapytania.

Wątek ten reprezentowany jest przez 2 wybrane prace: [SBCD09] oraz [SCW12].

W pracy [SBCD09] zaproponowano istotne ulepszenie podejścia rozważanego wcześniej w pracy innych badaczy [BCDG08], z którymi autor następnie nawiązał

współpracę naukową. To wcześniejsze podejście polegało na zastosowaniu zmodyfikowanego kombinatorycznego problemu pokrycia zbioru (ang. *red-blue set cover*) w ten sposób, aby w oparciu o logi zapytań dobrać taki zbiór podpowiedzi, który optymalizuje pewną funkcję celu. Funkcja ta modeluje m.in. dywersyfikację zbioru podpowiedzi a także trzy inne pożądane cechy wyniku. Ponieważ tak zdefiniowany problem jest NP-trudny, w pracy [BCDG08] zaproponowano heurystyczny algorytm zachłanny pozwalający na efektywne obliczenie rozwiązania przybliżonego.

Wkład autora stanowi opisane w [SBCD09] istotne rozwinięcie i ulepszenie podejścia opisanego w [BCDG08]. Polegało ono zarówno na ulepszeniu teoretycznego modelu problemu poprzez zauważenie pewnych wad uprzednio zdefiniowanej funkcji celu i zaproponowanie odpowiednich jej ulepszeń, jak i na zaproponowaniu nowej efektywnej metody rozwiązania tak zdefiniowanego problemu poprzez adaptację heurystyki *symulowanego wychładzania* [KGV+83] do obliczenia rozwiązania przybliżonego. Autor dokonał też całkowitej samodzielnej implementacji zaproponowanej ulepszonej funkcji celu oraz heurystycznego algorytmu i wykonał eksperymenty ewaluacyjne na unikatowych danych rzeczywistych (podzbiorze logów zapytań wyszukiwarki Yahoo!).

Praca [SBCD09] wykonywana była we współpracy z zespołem "Yahoo! Research", ośrodka badawczego jednej z dwóch największych na świecie komercyjnych wyszukiwarek internetowych, podczas pobytu autora na wewnętrznym krótkim stażu naukowo-badawczym na zaproszenie tego zespołu. Dzięki temu możliwe było wykonanie eksperymentów ewaluacyjnych na unikatowym cennym zbiorze danych przemysłowych. Dane takie, jako wrażliwe, są praktycznie niedostępne naukowcom poza kilkoma wyspecjalizowanymi firmami na świecie.

W związku z w/w niedostępnością publiczną danych, pomimo uzyskania zaprezentowanych w pracy [SBCD09] obiecujących wyników eksperymentalnych, które pokazały wyższość nowego podejścia nad proponowanym wcześniej w [BCDG08] i pomimo, że badany problem był bardzo interesujący naukowo, kontynuacja tego konkretnego projektu przez autora była niemożliwa po zakończeniu w/w stażu.

Wkład autora w publikację [SBCD09] jest kluczowy, gdyż obejmuje inspirację badań, większość koncepcji i ich pełną implementację wraz z kodem do ewaluacji oraz koordynację wszystkich prac. Pozostali autorzy uczestniczyli w dyskusjach koncepcyjnych, wstępnym przygotowaniu danych przemysłowych i niewielkim udziale w pisaniu artykułu w zakresie wstępu oraz prezentacji części wyników.

Prace przedstawione w publikacji [SBCD09] prezentowane były na specjalistycznym międzynarodowym warsztacie WSCD 2009 poświęconym analizie logów wyszukiwarek WWW w ramach konferencji WSDM 2009.

4.4.6 Generowanie zdywersyfikowanego zbioru podpowiedzi zapytań poprzez dyspersję (praca [SCW12])

Ostatnia pozycja [SCW12] omawianego cyklu wybranych prac dotyczy również problemu generowania zdywersyfikowanego zbioru podpowiedzi zapytań w wyszukiwarce internetowej w oparciu o zgromadzone uprzednio logi zapytań (podobnie jak w pracy [SBCD09] opisanego w sekcji 4.4.5).

W pracy [SCW12] stosuje się jednak zupełnie inne podejście niż w pracy [SBCD09]. Podejście to polega na adaptacji algorytmu MMR ("Maximal Marginal Relevance"),

zaproponowanego wcześniej w [CG98], do generowania zdywersyfikowanych wyników wyszukiwania i podsumowań tekstu. Podejście polega na zachłannym dobieraniu kolejnych elementów zbioru tak, aby były podobne do wyjściowego zapytania, ale jednocześnie niepodobne do innych wybranych podpowiedzi (a więc zdywersyfikowane). W pracy [SCW12] rozważane były różne warianty funkcji podobieństwa, m.in. bazująca na odległości edycyjnej Levenshteina oraz na podobieństwie semantycznym obliczonym za pomocą mapowania podpowiedzi na drzewo ontologii Wikipedii.

Z uwagi na niedostępność publiczną wysokiej jakości międzynarodowych logów wyszukiwarek (por. sekcja 4.4.5) nawiązano współpracę z wiodącą[4] krajową wyszukiwarką internetową `Netsprint.pl` i uzyskano tymczasowy dostęp do podzbioru rzeczywistych logów tej wyszukiwarki, aby wykonać tymczasową demonstrację działania zaproponowanych algorytmów. Uzyskano wyniki, które wskazywały, że dla wieloznacznych zapytań zaproponowane podejście jest w stanie wygenerować trafne i zarazem zdywersyfikowane podpowiedzi, które pokrywają wiele różnych aspektów oryginalnego zapytania.

Wkład autora w tej pracy jest większościowy i dotyczy zarówno sformułowania problemu jak propozycji modelu, obu przedstawionych algorytmów generowania podpowiedzi, oraz dwóch z trzech metod obliczania podobieństwa podpowiedzi.[5]

Opisane w pracy [SCW12] podejście znalazło praktyczne zastosowanie w działającym eksperymentalnym systemie wyszukiwawczym rozwijanym w IPI PAN. Dokładniej, stanowiło ono podstawę do systemu automatycznych podpowiedzi zapytań w największej eksperymentalnej *semantycznej* wyszukiwarce internetowej dla języka polskiego NEKST [nek]. Oprócz więc aspektów naukowych praca [SCW12] stanowi wkład w rozwój nowoczesnych narzędzi wyszukiwawczych działających w praktyce.

Wyniki prac związanych z publikacją [SCW12] przedstawiane były na międzynarodowej konferencji SIIS 2012.

Wniosek dotyczący tematyki badawczej związanej z pracą [SCW12] wygrał konkurs NCN na finansowanie grantem N N 516 481940, którego autor był kierownikiem.

4.5 Podsumowanie cyklu prac

Opisywana wyżej tematyka badawcza autora dotycząca wielu aspektów dywersyfikacji informacji wydaje się pionierska w skali kraju, w tym sensie, że autorowi nie są znane wcześniejsze prace krajowe w tej tematyce. Natomiast tematyka dywersyfikacji informacji spotyka się od niedawna z rosnącym zainteresowaniem międzynarodowej społeczności naukowej o czym świadczy np. regularne ostatnio organizowanie specjalistycznych warsztatów dotyczących dywersyfikacji informacji na uznanych międzynarodowych konferencjach branżowych (np. ACM WSDM, WWW, ACM CIKM, ACM KDD, IEEE/ACM/WIC WI) oraz wysoka liczba publikowanych przez innych badaczy prac o tej tematyce na takowych konferencjach międzynarodowych.

[4] W czasie prowadzenia przedmiotowych badań

[5] Metoda oparta na podobieństwie semantycznym jest autorstwa jednego ze współautorów

Liczba publikacji ogółem	65
(w tym) publikacje po doktoracie	58
Indeksowanych w Web of Science	19
Na liście JCR	4
Łączna liczba cytowań *):	393
H-indeks *):	11
Cytowania wg Web of Science	19
H-indeks wg Web of Science	3
Łączna liczba punktów wg MNISW **)	około 250

Tablica 1: Sumaryczny dorobek publikacyjny.
*) Do obliczenia łącznej liczby cytowań i wartości H-indeks korzystano z ogólnie dostępnych narzędzi w tym serwisu Google Scholar. W ten sposób obliczone cytowania mogą zawierać niewielką liczbę (do ok. 10%) autocytowań. Obliczeń dokonano na dzień 12.12.2015)
**) wg. punktacji ministerialnej zgodnie z datami publikacji

Uzyskane wyniki zaprezentowane w publikacjach stanowiących cykl prezentowany w niniejszym autoreferacie oraz ich cytowania na forum międzynarodowym w tym na konferencjach branżowych oraz w międzynarodowych czasopismach, czasami mimo niedługiego czasu od ukazania się niektórych z nich świadczą, że wyżej opisane prace stanowią zauważalny konkretny wkład do międzynarodowych badań nad tą nową dziedziną.

5 Omówienie powiązanych i pozostałych osiągnięć naukowo-badawczych.

Całość dorobku naukowo-badawczego autora wykracza daleko poza wyodrębniony cykl 6 publikacji stanowiący osiągniecie a dotyczący zagadnień dywersyfikacji. Łączny dorobek autora liczy 65 publikacji i cechuje go wysoka wszechstronność tematyki i różnorodność stosowanych technik. Tematyka dywersyfikacji informacji stanowi w nim nie więcej niż 30% i jest podejmowana najwcześniej od 2009 roku.

Pozostałe publikacje ukazujące się od roku 2003 (przy czym dorobek po doktoracie obejmuje prace opublikowanie po roku 2004) obejmują szeroką tematykę od zagadnień wyszukiwarek WWW i eksploracji danych, przez przetwarzanie języka naturalnego, uczenie maszynowe, sieci społeczne, semantyczne grafy wiedzy i elementy sztucznej inteligencji obliczeniowej.

Interdyscyplinarność badań wynika ze współpracy zarówno z przedstawicielami przemysłu jak i np. lingwistami komputerowymi i socjologami.

Tabela 1 prezentuje zbiorcze statystyki dotyczące dorobku publikacyjnego i informacje bibliometryczne.

Poniżej, wymieniono wybrane publikacje autora w formie umownego podziału na 6 punktów tematycznych (wraz z podpunktami) w kolejności wielkości dorobku. Podział ten nie jest jedynym możliwym z uwagi na przenikanie się tematyki i stosowanego aparatu w poszczególnych pracach (np. niektóre zagadnienia eksploracji danych przeplatają się z zagadnieniami uczenia maszynowego lub sieci

społecznych, etc.).

Punkt 1. ("zagadnienia i zastosowania dywersyfikacji informacji") w poniższym podziale obejmuje zbiór publikacji dotyczących różnych aspektów zagadnienia *dywersyfikacji* i jest bezpośrednio związany z tematyką osiągnięcia opisywanego w autoreferacie, ściślej, stanowi jego nadzbiór.

Punkt 2. ("eksploracja danych WWW") zawiera 4 prace na temat analizy grafu WWW i wariantów algorytmu PageRank z lat 2003-2004 opublikowane przed doktoratem, i samą rozprawę doktorską.

Wszystkie pozostałe prace w tym i pozostałych punktach zostały opublikowane po doktoracie.

Każda z wymienionych publikacji występuje w poniższym zestawieniu tylko jednokrotnie za wyjątkiem punktu 6., który powtarza 3 publikacje z punktu 1., ale uwypuklając użyte techniki sztucznej inteligencji.

W nawiasach okrągłych podano łączne liczby publikacji w każdym podpunkcie.

1. zagadnienia i zastosowania dywersyfikacji informacji (16):
    - zdywersyfikowane podsumowania encji w semantycznych grafach wiedzy (5) [SPS10a, Syd11, SPS11, KKRS13, SPS13]
    - zdywersyfikowane podpowiedzi zapytań w wyszukiwarkach WWW (3) [SBCD09, SCW12, SMN+15]
    - zdywersyfikowane wyszukiwanie encji podobnych do zadanych (1) [MSS14]
    - teoretyczne podstawy dywersyfikacji informacji (3) [Syd14c, Syd14b, Syd14a]
    - rola dywersyfikacji w kooperacyjnych sieciach społecznych (3) [SSC14, BJLJS14, BSSC16]
    - dywersyfikacja populacji w algorytmach genetycznych (1) [STS15]
2. eksploracja danych WWW (15) (ang. web mining)
    - algorytm PageRank i jego warianty (w tym rozprawa doktorska) (7) [Syd04b, Syd04c, Syd05c, Syd05e, Syd05d, Syd05a, Syd04a]
    - analiza grafu sieci WWW i ruchu w sieci WWW (4) [KS03, KS04, Syd05d, Syd05b]
    - systemy inteligentnego zbierania dokumentów WWW (2) [CSS07, KS13]
    - zwalczanie spamu wyszukiwarkowego (1) [SPWC08]
    - automatyczne podpowiedzi w wyszukiwarkach WWW w oparciu o analizę logów zapytań (1) [MNSS13]
3. przetwarzanie języka naturalnego (10)

- automatyczna lematyzacja i sugerowanie form podstawowych
  (6) [PS07b, PSK07, PS07a, PWPS08, PSW09, PWS09]
- ekstrakcja relacji z tekstów polskich
  (3) [WS12b, WS12a, MCSK13]
- zastosowania w wykrywaniu spamu WWW
  (1) [PSW08]

4. uczenie maszynowe i zastosowania w sieciach społecznych i WWW (9)

   - predykcja zaufania w sieciach społecznych
     (3) [Syd08, BSW09, BS10]
   - automatyczna klasyfikacja dokumentów tekstowych
     (2) [PS05, PS06]
   - analiza i predykcja zachowań użytkowników WWW
     (2) [DKS08, JS08]
   - grupowanie trajektorii obiektów ruchomych
     (1) [LDSMS14]
   - zagadnienie wielo-etykietowania
     (1) [ŁS13]

5. semantyczne grafy wiedzy (6) (prace inne niż wymienione w p.1)

   - rangowanie wyników wyszukiwania w semantycznych grafach wiedzy
     (1) [ERS+09]
   - wizualizacja graficznych podsumowań encji
     (1) [SPS12]
   - podsumowywanie encji (wersje wczesne, bez dywersyfikacji)
     (2) [SPSS10, SPS10b]
   - wyszukiwanie encji podobnych do zadanych (wersje bez dywersyfikacji)
     (2) [MSS13, SCSS15]

6. algorytmy sztucznej inteligencji w trudnych problemach optymalizacyjnych (występują też w p.1)

   - algorytm genetyczny z nowym dywersyfikującym operatorem selekcji
     (1) [STS15]
   - symulowane wychładzanie w generowaniu podpowiedzi zapytań
     (1) [SBCD09]
   - algorytm mrówkowy w generowaniu podsumowań encji
     (1) [KKRS13]

Dodatkowe informacje o dorobku umieszczone są w załączonym wykazie.

Literatura

[BCDG08] Francesco Bonchi, Carlos Castillo, Debora Donato, and Aristides Gionis. Topical query decomposition. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 52–60, New York, NY, USA, 2008. ACM.

[BJLJS14] Leszek Bukowski, Michał, Jankowski-Lorek, Szymon Jaroszewicz, and Marcin Sydow. What makes a good team of wikipedia editors? a preliminary statistical analysis. In Akiyo Nadamoto, Adam Jatowt, Adam Wierzbicki, and Jochen L. Leidner, editors, *Social Informatics*, volume 8359 of *Lecture Notes in Computer Science*, pages 14–28. Springer Berlin Heidelberg, 2014.

[BLY12] Allan Borodin, Hyun Chul Lee, and Yuli Ye. Max-sum diversification, monotone submodular functions and dynamic updates. In *Proceedings of the 31st Symposium on Principles of Database Systems*, PODS '12, pages 155–166, New York, NY, USA, 2012. ACM.

[BS10] Piotr Borzymek and Marcin Sydow. Trust and distrust prediction in social network with combined graphical and review-based attributes. In Piotr Jędrzejowicz, Ngoc Thanh Nguyen, Robert J. Howlett, and Lakhmi C. Jain, editors, *KES-AMSTA (1)*, volume 6070 of *Lecture Notes in Computer Science*, pages 122–131. Springer, 2010.

[BSSC16] Katarzyna Baraniak, Marcin Sydow, Jacek Szejda, and Dominika Czerniawska. Studying the role of diversity in open collaboration network: Experiments on wikipedia. In *Advances of Network Science (Proc. of the NetSci-X 2016 Conference)*, volume 9564 of *Lecture Notes in Computer Science*, chapter 8. Springer, 2016.

[BSW09] Piotr Borzymek, Marcin Sydow, and Adam Wierzbicki. Enriching trust prediction model in social network with user rating similarity. In Katarzyna Wegrzyn-Wolska Ajith Abraham, Vaclav Snasel, editor, *Proceedings of the 1st International Conference on Computational Aspects of Social Networks (CASoN 2009)*, pages 40–47, Los Alamitos, NY, USA, 2009. IEEE Computer Society.

[CG98] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, New York, NY, USA, 1998. ACM.

[CH96] Barun Chandra and Magnus M. Halldórsson. Facility dispersion and remote subgraphs. In Rolf Karlsson and Andrzej Lingas, editors, *Algorithm Theory — SWAT'96*, volume 1097 of *Lecture Notes in Computer Science*, pages 53–65. Springer Berlin Heidelberg, 1996.

[CSS07] C. Castillo, B. Starosta, and M. Sydow. Crawl.pl: Measuring statistical and structural properties of the polish web. *Studia Informatica*, 1(8):43–73, 2007.

[Dij59] E.W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[DKS08] Krzysztof Dembczyński, Wojciech Kotłowski, and Marcin Sydow. Effective prediction of web user behaviour with user-level models. *Fundam. Inf.*, 89(2-3):189–206, April 2008.

[ERS+09] Shady Elbassuoni, Maya Ramanath, Ralf Schenkel, Marcin Sydow, and Gerhard Weikum. Language-model-based ranking for queries on rdf-graphs. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 977–986, New York, NY, USA, 2009. ACM.

[GS09] Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach for result diversification. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 381–390, New York, NY, USA, 2009. ACM.

[JS08] Joanna Jaworska and Marcin Sydow. Behavioural targeting in on-line advertising: An empirical study. In *WISE '08: Proceedings of the 9th international conference on Web Information Systems Engineering*, pages 62–76, Berlin, Heidelberg, 2008. Springer-Verlag.

[KGV+83] Scott Kirkpatrick, C Daniel Gelatt, Mario P Vecchi, et al. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[KKRS13] Witold Kosiński, Tomasz Kuśmierczyk, Paweł Rembelski, and Marcin Sydow. Application of ant-colony optimisation to compute diversified entity summarisation on semantic knowledge graphs. In *Proc. of International IEEE AAIA 2013/FedCSIS Conference, Annals of Computer Science and Information Systems*, volume 1, pages 69–76, 2013.

[KS03] Mieczyslaw A. Kłopotek and Marcin Sydow. Uncorrelating pagerank and in-degree in a synthetic web model. In Adnan Yazici and Cevat Sener, editors, *Computer and Information Sciences - ISCIS 2003*, volume 2869 of *Lecture Notes in Computer Science*, pages 139–146. Springer Berlin Heidelberg, 2003.

[KS04] Mieczyslaw A. Kłopotek and Marcin Sydow. Towards a more realistic web graph model. In Mieczyslaw A. Kłopotek, SlawomirT. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, volume 25 of *Advances in Soft Computing*, pages 321–330. Springer Berlin Heidelberg, 2004.

[KS13] Tomasz Kuśmierczyk and Marcin Sydow. Towards a keyword-focused web crawler. In Mieczysław A. Kłopotek, Jacek Koronacki, Małgorzata Marciniak, Agnieszka Mykowiecka, and SławomirT. Wierz-

choń, editors, *Language Processing and Intelligent Information Systems*, volume 7912 of *Lecture Notes in Computer Science*, pages 187–197. Springer Berlin Heidelberg, 2013.

[KSI+08] Gjergji Kasneci, Fabian M. Suchanek, Georgiana Ifrim, Maya Ramanath, and Gerhard Weikum. NAGA: Searching and Ranking Knowledge. In *24th International Conference on Data Engineering (ICDE 2008)*. IEEE, 2008.

[LDSMS14] Bo Liu, E.N. De Souza, S. Matwin, and M. Sydow. Knowledge-based clustering of ship trajectories using density-based approach. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 603–608. IEEE, Oct 2014.

[ŁS13] Michał, Łukasik and Marcin Sydow. Threshold ml-knn: Statistical evaluation on multiple benchmarks. In Mieczysław A. Kłopotek, Jacek Koronacki, Małgorzata Marciniak, Agnieszka Mykowiecka, and SławomirT. Wierzchoń, editors, *Language Processing and Intelligent Information Systems*, volume 7912 of *Lecture Notes in Computer Science*, pages 198–205. Springer Berlin Heidelberg, 2013.

[MCSK13] Marcin Mirończuk, Dariusz Czerski, Marcin Sydow, and Mieczysław A. Kłopotek. Language-independent information extraction based on formal concept analysis. In *Informatics and Applications (ICIA),2013 Second International Conference on*, pages 323–329, 2013.

[MNSS13] Cristina Ioana Muntean, Franco Maria Nardini, Fabrizio Silvestri, and Marcin Sydow. Learning to shorten query sessions. In *Proceedings of the 22nd international conference on World Wide Web companion*, WWW '13 Companion, pages 131–132, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[MSS13] Steffen Metzger, Ralf Schenkel, and Marcin Sydow. Qbees: query by entity examples. In *Proceedings of the 22nd ACM international conference on Conference on Information & Knowledge Management*, CIKM '13, pages 1829–1832, New York, NY, USA, 2013. ACM.

[MSS14] Steffen Metzger, Ralf Schenkel, and Marcin Sydow. Aspect-based similar entity search in semantic knowledge graphs with diversity-awareness and relaxation. In *Proc. of the 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2014*, pages 60–69, 2014.

[mtu] http://www.mturk.com.

[nek] http://nekst.pl.

[PS05] J. Piskorski and M. Sydow. Experiments on classification of polish newspaper articles. *Archives of Control Sciences*, Vol. 15, no. 4:625–636, 2005.

[PS06] J. Piskorski and M. Sydow. Fine-tuning n-gram-based text classifier for highly inflective languages. In *Challenging Problems of Computer Science, Artificial Intelligence and Soft Computing*, pages 494 – 499. Academic Publishing House EXIT, Polish Neural Society, IEEE Computational Intelligence Society - Poland Chapter, 2006.

[PS07a] Jakub Piskorski and Marcin Sydow. String distance metrics for reference matching and search query correction. In Witold Abramowicz, editor, *Business Information Systems*, volume 4439 of *Lecture Notes in Computer Science*, pages 353–365. Springer Berlin / Heidelberg, 2007. 10.1007/978-3-540-72035-5.27.

[PS07b] Jakub Piskorski and Marcin Sydow. Usability of string distance metrics for name matching tasks in polish. In *Human Language Technologies as a Challenge for Computer Science and Linguistics, Proc. of LTC'07*, pages 403–407. Wydawnictwo Poznańskie Sp. z o.o., 2007.

[PSK07] Jakub Piskorski, Marcin Sydow, and Anna Kupść. Lemmatization of polish person names. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, ACL '07, pages 27–34, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[PSW08] Jakub Piskorski, Marcin Sydow, and Dawid Weiss. Exploring linguistic features for web spam detection: a preliminary study. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 25–28, New York, NY, USA, 2008. ACM.

[PSW09] Jakub Piskorski, Marcin Sydow, and Karol Wieloch. Comparison of string distance metrics for lemmatisation of named entities in polish. pages 413–427, 2009.

[PWPS08] Jakub Piskorski, Karol Wieloch, M Pikula, and Marcin Sydow. Towards person name matching for inflective languages. In *WWW 2008 Workshop NLP Challenges in the Information Explosion Era*, 2008.

[PWS09] Jakub Piskorski, Karol Wieloch, and Marcin Sydow. On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages. *Information Retrieval*, 12(3):275–299, 2009.

[SBCD09] Marcin Sydow, Francesco Bonchi, Carlos Castillo, and Debora Donato. Optimising topical query decomposition. In *Proceedings of the 2009 Workshop on Web Search Click Data*, WSCD '09, pages 43–47, New York, NY, USA, 2009. ACM.

[SCSS15] Grzegorz Sobczak, Mateusz Chochół, Ralf Schenkel, and Marcin Sydow. iQbees: Towards interactive semantic entity search based on maximal aspects. In Floriana Esposito, Olivier Pivert, Mohand-Said Hacid, Zbigniew Raś, and Stefano Ferilli, editors, *Foundations of Intelligent Systems*, volume 9384 of *Lecture Notes in Computer Science*,

pages 259–264. Springer International Publishing, 2015. 10.1007/978-3-319-25252-0-28.

[SCW12] Marcin Sydow, Krzysztof Ciesielski, and Jakub Wajda. Introducing diversity to log-based query suggestions to deal with underspecified user queries. In Pascal Bouvry, Mieczyslaw Klopotek, Franck Leprévost, Malgorzata Marciniak, Agnieszka Mykowiecka, and Henryk Rybinski, editors, *Security and Intelligent Information Systems*, volume 7053 of *Lecture Notes in Computer Science*, pages 251–264. Springer Berlin / Heidelberg, 2012. 10.1007/978-3-642-25261-7-20.

[SMN+15] Marcin Sydow, Cristina Ioana Muntean, Franco Maria Nardini, Stan Matwin, and Fabrizio Silvestri. MUSETS: Diversity-aware web query suggestions for shortening user sessions. In Floriana Esposito, Olivier Pivert, Mohand-Said Hacid, Zbigniew Ras, and Stefano Ferilli, editors, *Foundations of Intelligent Systems*, volume 9384 of *Lecture Notes in Computer Science*, pages 237–247. Springer International Publishing, 2015. 10.1007/978-3-319-25252-0-26.

[SPS10a] Marcin Sydow, Mariusz Pikuła, and Ralf Schenkel. DIVERSUM: Towards diversified summarisation of entities in knowledge graphs. In *Proceedings of Data Engineering Workshops (ICDEW) at IEEE 26th ICDE Conference*, pages 221–226. IEEE, 2010.

[SPS10b] Marcin Sydow, Mariusz Pikuła, and Ralf Schenkel. Entity summarization with limited edge budget on undirected and directed knowledge graphs. *Investigationes Linguisticae*, 21:76–89, 2010.

[SPS11] Marcin Sydow, Mariusz Pikuła, and Ralf Schenkel. To diversify or not to diversify entity summaries on rdf knowledge graphs? In Marzena Kryszkiewicz, Henryk Rybiński, Andrzej Skowron, and Zbigniew Ras, editors, *Foundations of Intelligent Systems, Proc. of the 19th ISMIS Conference, Warsaw, Poland, 2011*, volume 6804 of *Lecture Notes in Artificial Intelligence*, pages 490–500. Springer Berlin / Heidelberg, 2011. 10.1007/978-3-642-21916-0-53.

[SPS12] Grzegorz Sobczak, Mariusz Pikuła, and Marcin Sydow. Agnes: A novel algorithm for visualising diversified graphical entity summarisations on knowledge graphs. In Li Chen, Alexander Felfernig, Jiming Liu, and Zbigniew Raś, editors, *Foundations of Intelligent Systems*, volume 7661 of *Lecture Notes in Computer Science*, pages 182–191. Springer Berlin Heidelberg, 2012.

[SPS13] Marcin Sydow, Mariusz Pikuła, and Ralf Schenkel. The notion of diversity in graphical entity summarisation on semantic knowledge graphs. *Journal of Intelligent Information Systems*, 41:109–149, 2013.

[SPSS10] Marcin Sydow, Mariusz Pikuła, Ralf Schenkel, and Adam Siemion. Entity summarisation with limited edge budget on knowledge graphs. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, pages 513–516. IEEE, 2010.

[SPWC08] Marcin Sydow, Jakub Piskorski, Dawid Weiss, and Carlos Castillo. *Fighting Web Spam*, volume 19 of *NATO Science for Peace and Security Series D: Information and Communication Security*, pages 134–153. IOS Press, 2008.

[SSC14] Jacek Szejda, Marcin Sydow, and Dominika Czerniawska. Does a 'renaissance man' create good wikipedia articles? In *International Conference on Knowledge Discovery and Information Retrieval (KDIR 2014)*, page 425, 2014.

[STS15] Anna Strzeżek, Ludwik Trammer, and Marcin Sydow. Divergene: Experiments on controlling population diversity in genetic algorithm with a dispersion operator. In *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems*, volume 5 of *Annals of Computer Science and Information Systems*, pages 155–162, Sept 2015.

[Syd04a] M. Sydow. Link analysis of the web graph. measurements, models and algorithms for web information retrieval (phd thesis), 2004.

[Syd04b] Marcin Sydow. Extensions of pagerank. the rbs algorithm. In Mieczysław A. Kłopotek, SławomirT. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, volume 25 of *Advances in Soft Computing*, pages 389–396. Springer Berlin Heidelberg, 2004.

[Syd04c] Marcin Sydow. Random surfer with back step. In *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers &Amp; Posters*, WWW Alt. '04, pages 352–353, New York, NY, USA, 2004. ACM.

[Syd05a] M. Sydow. Random surfer with back step. *Fundamenta Informaticae*, Vol. 68, nr 4:379–398, 2005.

[Syd05b] M. Sydow. Studying dependencies among web traffic and link analysis data using perceptron. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pages 124–127, Sept 2005.

[Syd05c] Marcin Sydow. Approximation quality of the rbs ranking algorithm. In Mieczysław A. Kłopotek, SławomirT. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, volume 31 of *Advances in Soft Computing*, pages 289–296. Springer Berlin Heidelberg, 2005.

[Syd05d] Marcin Sydow. Can link analysis tell us about web traffic? In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, WWW '05, pages 954–955, New York, NY, USA, 2005. ACM.

[Syd05e] Marcin Sydow. Can one out-link change your pagerank? In Piotr S. Szczepaniak, Janusz Kacprzyk, and Adam Niewiadomski, editors, *Advances in Web Intelligence*, volume 3528 of *Lecture Notes in Computer Science*, pages 408–414. Springer Berlin Heidelberg, 2005.

[Syd08] M. Sydow. Towards using contextual information to learn trust metric in social networks: A proposal. In *Proceedings of the 2nd International Workshop on Combining Context with Trust, Security and Privacy, in conjunction with IFIPTM08*, 2008.

[Syd11] Marcin Sydow. Towards the foundations of diversity-aware node summarisation on knowledge graphs. In *Proceedings of "Diversity in Document Retrieval" Workshop, European Conference on Information Retrieval*. Springer, 2011.

[Syd14a] Marcin Sydow. Approximation guarantees for max sum and max min facility dispersion with parameterised triangle inequality and applications in result diversification. *Mathematica Applicanda*, 42(2):241–257, 2014.

[Syd14b] Marcin Sydow. Improved approximation guarantee for max sum diversification with parameterised triangle inequality. In Troels Andreasen, Henning Christiansen, Juan-Carlos Cubero, and Zbigniew. Ras, editors, *Foundations of Intelligent Systems*, volume 8502 of *Lecture Notes in Computer Science*, pages 554–559. Springer International Publishing, 2014.

[Syd14c] Marcin Sydow. Towards integrity in diversity-aware small set selection and visualisation tasks. pages 480–484. SCITEPRESS, 2014.

[wik] http://wikipedia.org.

[WS12a] Alina Wróblewska and Marcin Sydow. Debora: Dependency-based method for extracting entity-relationship triples from open-domain texts in polish. In Li Chen, Alexander Felfernig, Jiming Liu, and Zbigniew W. Raś, editors, *Foundations of Intelligent Systems*, volume 7661 of *Lecture Notes in Computer Science*, pages 155–161. Springer Berlin Heidelberg, 2012.

[WS12b] Alina Wróblewska and Marcin Sydow. Dependency-based extraction of entity-relationship triples from polish open-domain texts. In *Proc. of Artificial Intelligence Studies*, volume 7(30), pages 61–70. Publ. House of University of Natural Sciences and Humanities, Siedlce, 2012.

[yag] http://mpi-inf.mpg.de/yago.