STUDIA OECONOMICA POSNANIENSIA 2013, vol. 1, no. 2(251)

Jacek Małyszko
Uniwersytet Ekonomiczny w Poznaniu, Wydział Informatyki i Gospodarki Elektronicznej, Katedra Informatyki Ekonomicznej
jacek.malyszko@kie.ue.poznan.pl

UJEDNOZNACZNIANIE POJĘĆ DLA JĘZYKA POLSKIEGO DLA POTRZEB BUDOWANIA TOŻSAMOŚCI UŻYTKOWNIKÓW INTERNETU

Streszczenie: Wraz z przenoszeniem się aktywności ludzi do Internetu, coraz istotniejsze staje się zagadnienie wirtualnych tożsamości użytkownika, rozumianych jako przetwarzalne reprezentacje cech danego użytkownika sieci. Tożsamości takie mają umożliwić powiadamianie systemów, z którymi użytkownik wchodzi w interakcje, o istotnych jego cechach, na przykład dla potrzeb personalizacji treści dostarczanej przez system. Opracowanie reprezentacji różnych cech, która miałaby być zrozumiała dla wielu niepowiązanych ze sobą systemów, jest trudne. Dodatkowym wyzwaniem jest opracowanie sposobów pozyskiwania informacji o charakterystykach użytkownika służących do utworzenia jego tożsamości. Artykuł ma na celu przeanalizowanie dwóch zagadnień, które mogą znaleźć zastosowanie w rozwiązaniu omawianych problemów. Są nimi: semantyczne modelowanie użytkowników, które może pozwolić na konstruowanie modeli możliwych do wykorzystania w wielu różnych systemach, oraz ujednoznacznianie pojęć wyekstrahowanych z czytanych przez użytkownika artykułów, będące jednym ze sposobów pozyskiwania informacji o jego potrzebach. Na podstawie przeprowadzonej analizy podjęta została próba określenia kierunków badań w omawianych zakresach.

Słowa kluczowe: semantyczne modelowanie użytkowników, zarządzanie tożsamością, ujednoznacznianie pojęć.

Klasyfikacja JEL: C88, D83.

Wstęp

Od kilku lat w dziedzinie informatyki można zaobserwować coraz większe zainteresowanie tematem tożsamości użytkowników Internetu. Tożsamości można rozumieć z jednej strony w kontekście mechanizmów umożliwiających potwierdzanie tożsamości danej osoby w różnego rodzaju systemach i podczas przeprowadzania transakcji internetowych, a z drugiej w odniesieniu do reprezentacji cech danego użytkownika w sieci [Nabeth i Hildebrandt 2005]. Jednocześnie istotą modelowania tożsamości internetowych jest bardziej całościowe ujęcie tych aspektów, niż miało

to miejsce w dotychczasowych rozwiązaniach. Przykładem obszaru, w którym wyraźnie widać takie podejście, jest modelowanie użytkowników. Tradycyjnie rozumiany model użytkownika ma za zadanie odzwierciedlać jedynie pewien fragment charakterystyki użytkownika, odpowiadający potrzebom pojedynczego systemu, w którym dany model jest przechowywany i wykorzystywany. Na przykład, jeśli model utrzymywany jest na potrzeby systemu rekomendującego utwory muzyczne, są w nim przechowywane reprezentacje gustów muzycznych danego użytkownika. Natomiast tożsamość użytkownika ma za zadanie przechowywać globalny model użytkownika, reprezentujący szersze spektrum jego cech i potrzeb związanych z całą jego działalnością w sieci.

Temat globalnego modelowania użytkownika jest jednym z obszarów zainteresowania projektu Ego – Virtual Identity[1]. Celem projektu Ego jest utworzenie modelu oraz mechanizmów zarządzających wirtualną tożsamością użytkownika, rozumianą jako ewoluujący w czasie model użytkownika, a następnie zbudowanie prototypu systemu wspierającego ewolucję tożsamości i jej wykorzystanie przez systemy informatyczne.

Skonstruowanie systemu zarządzania wspomnianymi globalnymi modelami użytkowników niesie ze sobą wiele wyzwań. Można zdefiniować trzy główne pytania badawcze:

1. Jaka technika reprezentacji wiedzy powinna zostać zastosowana, aby możliwe było przechowywanie informacji o jak najszerszym zakresie charakterystyk użytkownika?
2. W jaki sposób należy zapewnić możliwość wykorzystania modelu w wielu niezależnych systemach?
3. W jaki sposób należy pozyskiwać informacje o charakterystykach użytkowników?

Artykuł ma na celu przeanalizowanie dotychczasowych prac w tym obszarze, a także wypracowanie wizji i rekomendacji dotyczących omawianych kwestii.

Szanse na rozwiązanie dwóch pierwszych problemów badawczych niesie ze sobą wykorzystanie semantycznych modeli użytkowników, czyli takich modeli, w których charakterystyki użytkowników są opisane za pomocą pojęć pochodzących z pewnej ontologii. Jednak założenia, że w tożsamości ma być przechowywany jak najszerszy zakres charakterystyk użytkownika oraz że tożsamość może być wykorzystywana przez dowolne systemy w sieci Internet, wprowadzają do tego zagadnienia nowe wyzwania. Sekcja 1 zawiera analizę literatury dla tego zagadnienia.

Jednym z możliwych rozwiązań trzeciego problemu, czyli pozyskiwania informacji o charakterystykach użytkowników, jest monitorowanie (oczywiście za zgodą użytkownika) dokumentów tekstowych, które użytkownik czyta w Internecie, w celu

[1] Więcej informacji na temat projektu na stronie http://kie.ue.poznan.pl/pl/project/ego-virtual-identity [dostęp: 18.01.2013].

automatycznej analizy ich tematyki i określenia na tej podstawie zainteresowań użytkownika. Ze względu na to, że tematyka artykułów z założenia może być dowolna, kluczową kwestią staje się ujednoznacznienie pojęć występujących w tych dokumentach. Analiza stanu badań z tego zakresu umieszczona jest w sekcji 2.

W sekcji 3 artykułu przedstawiono wnioski wypływające z analizy obecnego stanu wiedzy w omawianych dziedzinach w postaci wizji ich zastosowania dla potrzeb zarządzania internetowymi tożsamościami użytkowników. W ostatniej sekcji zawarto podsumowanie artykułu oraz zaprezentowano kierunki dalszych badań.

1. Semantyczne modelowanie użytkowników

Otwierając strony internetowe, użytkownicy mają określony cel: na przykład, przeglądając stronę z wiadomościami sportowymi, chcemy się dowiedzieć, jakie są najnowsze wyniki ligi piłki siatkowej (naszym celem jest zaspokojenie określonej potrzeby informacyjnej), odwiedzając sklep internetowy sprzedający płyty, chcemy kupić płytę z określonym, ulubionym przez nas gatunkiem (planujemy dokonać zakupu). Często nasze potrzeby są stałe: posługując się przytoczonym powyżej przykładem, mamy określony ulubiony gatunek muzyczny i interesują nas tylko płyty artystów reprezentujących właśnie ten nurt. Systemy informatyczne starają się rozpoznać stałe cechy użytkowników (jak ulubiona dyscyplina sportowa itp.), tak aby na przykład zawsze, gdy określony użytkownik odwiedzi daną stronę, na jej górze pojawiały się wiadomości dotyczące wyników meczy siatkówki, a nie piłki nożnej, której użytkownik nie lubi. Celem takiego działania jest spełnienie oczekiwań użytkownika i zachęcenie go do kolejnych odwiedzin lub zaprezentowanie użytkownikowi takich treści, które skłonią go do podjęcia pożądanej przez system akcji, np. zakupu. Tak rozpoznane potrzeby zapisuje się w systemach informatycznych w postaci tzw. modeli użytkowników. W tej sekcji zajmiemy się takimi właśnie modelami, konstruowanymi z wykorzystaniem ontologii.

1.1. Potrzeba ponownego wykorzystania modeli użytkowników

Model użytkownika definiuje się jako reprezentację informacji o danym użytkowniku, która jest niezbędna dla systemów adaptacyjnych do przeprowadzenia tzw. efektu adaptacji, czyli zapewnienia, że system zachowuje się w różny sposób dla różnych użytkowników [Brusilovsky i Millán 2007]. Systemy informatyczne przechowują w modelach takie informacje o użytkownikach, jak ich charakterystyki demograficzne czy geograficzne, listę obszarów ich wiedzy i kompetencji, ich gusta czy zainteresowania [Nabeth i Gasson 2005]. Wspomniany efekt adaptacji polega na modyfikacji zawartości lub funkcji danego systemu informatycznego w celu lepszego zaspokojenia potrzeb użytkownika [Brusilovsky i Millán 2007]. Z pojęciem systemów adaptacyjnych bardzo blisko powiązana jest również personalizacja, czyli zdolność

systemów do dostarczenia zawartości i usług dopasowanych do konkretnej osoby na podstawie znajomości jej preferencji i zachowań [Adomavicius i Tuzhilin 2005]. Właśnie z punktu widzenia personalizacji przeprowadzona zostanie dalsza analiza.

Tradycyjnie model użytkownika był przeznaczony do wykorzystania w konkretnym celu w pojedynczym systemie. Informacje o użytkowniku były przechowywane na przykład w postaci listy słów kluczowych lub wektora wskazującego stopień zainteresowania użytkownika określonymi tematami [Salton, Wong i Yang 1975; Sosnovsky i Dicheva 2010]. Pod koniec lat 90. XX wieku pojawiły się pierwsze głosy postulujące konieczność takiego konstruowania modeli, aby możliwe było wykorzystanie ich w wielu różnych systemach. Uniknięto by w ten sposób uciążliwej konieczności ponownego konstruowania modelu od podstaw w każdym kolejnym systemie, z którego użytkownik zaczyna korzystać. Umożliwiłoby to użytkownikowi natychmiastowe czerpanie korzyści z personalizacji również wówczas, gdy dopiero zaczyna korzystać z danego systemu [Kay 1999; Niederée i in. 2004].

Znaczna część prób rozwiązania problemu braku możliwości wykorzystania modeli użytkowników w wielu różnych systemach skupiła się na wykorzystaniu ontologii do opisywania charakterystyk użytkowników. Ontologia pierwotnie została zdefiniowana jako specyfikacja konceptualizacji określonej dziedziny [Gruber 1993]. Ontologia definiuje pojęcia i relacje między nimi, a także inne charakterystyki istotne z punktu widzenia modelowanej domeny. Na przykład, cechy użytkownika mogą być opisywane za pomocą trzech elementów: czasownika posiłkowego (ang. *auxiliary*), predykatu (ang. *predicate*) oraz zakresu wartości (ang. *range*). W przypadku zainteresowania danego użytkownika piłką siatkową, reprezentacja tego faktu za pomocą ontologii mogłaby przybrać następującą formę: *auxiliary = has interest, predicate = volleyball, range = low-medium-high*, gdzie wartości przyjmowane przez każdy z tych elementów są ustalone w wykorzystywanej ontologii [Heckmann 2005]. Modele użytkowników skonstruowane z wykorzystaniem ontologii w dalszej części pracy nazywane będą modelami semantycznymi.

W pracy L. Razmerity, A. Angehrna i A. Maedche'a [2003] wymieniono trzy różne wykorzystane ontologie w procesie modelowania użytkowników:

– ontologię użytkownika, która opisuje charakterystyki użytkownika podlegające modelowaniu; przykładem takiej ontologii jest GUMO, omówiona w sekcji 1.2,
– ontologię dziedzinową, definiującą domenę, w której stosowany będzie model; za jej pomocą można na przykład reprezentować zainteresowanie użytkownika danym tematem,
– ontologię dziennika zdarzeń (ang. *Log Ontology*), która definiuje znaczenie akcji podejmowanych przez użytkownika podczas jego interakcji z systemem; umożliwia ona automatyczne konstruowanie modelu użytkownika.

Konstruowanie modeli za pomocą ontologii może umożliwić ich wykorzystanie w wielu systemach. Dzieje się tak przede wszystkim dlatego, że z definicji ontologia jest wysoce sformalizowanym opisem pewnej domeny czy też reprezentacją wiedzy o danej domenie. Ontologia, którą wykorzystano do budowania modeli, stanowi więc

pewnego rodzaju dokumentację ułatwiającą ponowne ich wykorzystanie. Na ten fakt zwrócono uwagę w pracy J. Bourdeau i R. Mizoguchiego [2000], gdzie stwierdzono, że wykorzystanie ontologii w znacznym stopniu ułatwia późniejsze wykorzystanie wszelkich systemów sztucznej inteligencji oraz późniejsze wprowadzanie w nich zmian. Należy również zwrócić uwagę, że oprócz możliwości ponownego wykorzystania modeli w innych systemach, wzbogacenie modelowania użytkowników o semantykę przynosi również inne korzyści w postaci możliwości przeprowadzania dodatkowych wnioskowań, które mogą poprawić jakość personalizacji, co wykazano przykładowo w pracy S.S. Ananda, P. Kearney i M. Shapcotta [2007].

1.2. Podejścia do semantycznego modelowania użytkowników

Badania nad semantycznym modelowaniem użytkowników trwają od ponad 10 lat. Przez ten czas wypracowano wiele podejść do wykorzystania ontologii w tym procesie. Aktualny przegląd znacznej części badań w tym zakresie przedstawili S. Sosnovsky i D. Dicheva [2010]. Na potrzeby artykułu, zastosowanie modelowania użytkowników z wykorzystaniem ontologii można podzielić według kilku kryteriów, które zaprezentowano w tabeli.

Trzy wymiary zastosowania ontologii w modelowaniu użytkowników

Kryterium podziału	Podejścia
Obszar wykorzystania modelu	– pojedynczy system; wykorzystana ontologia dziedzinowa specyficzna dla tematyki systemu – brak ograniczeń, potencjalnie cały Internet; konieczność reprezentacji szerokiego zakresu charakterystyk użytkownika
Istniejąca / nowa ontologia	– wykorzystanie istniejącej ontologii – utworzenie ontologii specjalnie dla potrzeb modelowania
Ustalona / dynamiczna struktura modelu	– model użytkownika jako instancja ontologii, często z przypisanymi wartościami liczbowymi reprezentującymi na przykład zainteresowanie użytkownika danym tematem – struktura modelu tworzona (częściowo lub w całości) dynamicznie, na przykład z wykorzystaniem nauki ontologii (ang. *ontology learning*); modele różnych użytkowników mogą się od siebie różnić strukturą

Jedną z pierwszych prób wzbogacenia modelowania użytkowników o semantykę przeprowadzono podczas prac nad projektem SiteIF na Uniwersytecie w Trydencie [Stefani i Strappavara 1998]. Jest to przykład wykorzystania ontologii leksykalnej w procesie modelowania użytkowników. Wykorzystano tutaj WordNet[2] (opisany dokładniej w sekcji 2.1) i automatyczną naukę modelu na podstawie analizy treści stron

[2] Więcej informacji na temat WordNetu na stronie http://wordnet.princeton.edu/ [dostęp: 18.01.2013].

czytanych przez użytkownika w ramach danego portalu. Modelem użytkownika jest sieć synsetów (czyli pojęć w WordNecie odpowiadających znaczeniom wyrazów) wraz z przypisanymi wagami, obrazującymi stopień zainteresowania użytkownika danym tematem. Krawędzie między synsetami budowane są na podstawie analizy ich współwystępowania w dokumentach czytanych przez użytkowników. Zarówno węzłom, jak i krawędziom nadawane są wagi, które z upływem czasu ewoluują, odzwierciedlając zmiany w wykrytych zainteresowaniach użytkownika [Magnini i Strapparava 2001]. Podejście przyjęte w badaniach wydaje się obiecujące z punktu widzenia wykorzystania dla potrzeb budowania tożsamości. Struktura modeli jest zrozumiała i łatwa w przetwarzaniu, a WordNet jest ogólnie przyjętym standardem reprezentacji wiedzy leksykalnej. W związku z tym teoretycznie dowolny system może zawierać moduł personalizujący treść na podstawie modeli skonstruowanych dla potrzeb projektu SiteIF. Należy jednak zwrócić uwagę, że – jak już wspomniano – w omawianej pracy celem zastosowania ontologii w modelowaniu było jedynie usprawnienie procesu wnioskowania dla potrzeb ustalania rekomendacji.

Badania podobne do opisanych wyżej przeprowadzano jeszcze kilkakrotnie, nie wprowadzając jednak znaczących zmian w podejściu do rozwiązania problemu. Przykładem może być praca H. Zhanga, Y. Songa i H. Songa [2007]. Zaproponowany model ponownie składa się z pojęć, tym razem pochodzących z przyjętej ontologii dziedzinowej danego systemu, oraz z połączeń między nimi (połączenia takie charakteryzowane są tu przez kilka parametrów). Konstruowanie modelu odbywa się na drodze analizy i semantycznego wzbogacania logów z akcji użytkownika w trakcie sesji użytkowania systemu. Z każdej sesji tworzona jest sieć semantyczna reprezentująca sesję, która następnie jest łączona z siecią będącą modelem użytkownika.

Jako prostszy przykład wykorzystania ontologii do modelowania w celu ustalenia trafniejszych rekomendacji w pojedynczym systemie może również posłużyć praca A. Sieg, B. Mobashera i R. Burke'a [2010], gdzie modele użytkowników są instancjami ontologii dziedzinowej (rozumianej jako taksonomia, czyli hierarchia pojęć) z przypisanymi do pojęć wagami.

Inną motywacją (bliższą motywacji autora artykułu) jest wykorzystanie ontologii w celu umożliwienia ponownego wykorzystania modeli w różnych systemach. C. Niederée i in. [2004] skupili się na tworzeniu modelu kontekstu użytkownika (ang. *Unified User Context Model*, UUCM), który mógłby być wykorzystany przez różne systemy, z którymi użytkownik wchodzi w interakcję podczas wykonywania zadań mających wspólny cel, na przykład przygotowanie zajęć dla studentów bądź planowanie wakacji. Modelowaniu podlegają tu cztery wymiary:

- wzorzec kognitywny użytkownika (tradycyjne dla modelowania charakterystyki, takie jak obszary zainteresowania),
- aktualnie wykonywane zadanie i rola użytkownika w tym zadaniu,
- relacje między użytkownikiem a bytami w jego otoczeniu,
- środowisko użytkownika, czyli tradycyjne parametry dotyczące kontekstu użytkownika, jak czas, miejsce, urządzenie, język.

Użytkownik może być opisany w każdym z tych wymiarów za pomocą tak zwanych aspektów, które mogą przyjmować różne wartości. Nazwy aspektów i ich wartości pochodzą z przyjętych ontologii, przy czym należy zwrócić uwagę, że możliwe jest wykorzystanie w tym celu dowolnych ontologii. W omawianym artykule autorzy dodatkowo zaprezentowali scenariusz wykorzystania tak ustrukturyzowanego modelu, w którym użytkownik prezentuje systemom, z którymi wchodzi w interakcję, tak zwany paszport kontekstu (ang. *Passport Context*). Na jego podstawie systemy współdzielące z modelem te same ontologie są w stanie poznać potrzeby użytkownika i wykorzystać tę znajomość dla potrzeb personalizacji.

Odmienne cele przyjął w swojej pracy D. Heckmann [2005]. Podjął próbę utworzenia od podstaw ontologii mającej opisywać użytkownika, która następnie mogłaby zostać przyjęta przez twórców różnych systemów, co umożliwiłoby wymianę modeli pomiędzy takimi systemami. Ontologii tej nadano nazwę GUMO (ang. *General User Model Ontology*). Na ontologię składają się grupy wymiarów, w których modelowany jest użytkownik (np. MentalState, Demographics, Personality), wymiary wchodzące w skład tych grup, a także lista kategorii możliwych zainteresowań użytkowników. Inną próbę utworzenia uniwersalnej ontologii dla modelowania użytkowników opisali w artykule F. Cretton i A. La Calvé [2008]. Zaprezentowali badania mające na celu utworzenie uniwersalnego modelu użytkownika, tutaj zwanego GenOUM (ang. *Generic Ontology-based User Model*).

Przedstawiona analiza wskazuje, że przeprowadzono już wiele badań w kierunku opracowania modelu użytkownika, który możliwy byłby do wykorzystania w wielu systemach. Obszar ten wciąż jednak dynamicznie się rozwija i ciągle toczą się w nim prace [Sosnovsky i Dicheva 2010]. Wnioski z przeprowadzonej analizy płynące dla wymagań dla projektu Ego zostaną przedstawione w sekcji 3.1 artykułu.

2. Ujednoznacznianie pojęć

W poprzedniej sekcji przedstawiono zagadnienie modelowania użytkowników jako reprezentacji w systemie informatycznym informacji o cechach użytkowników. Wcześniej jednak trzeba zadbać o to, w jaki sposób takie informacje o użytkowniku można zdobyć. W przypadku globalnego modelu użytkownika, który ma przedstawiać jego charakterystykę w jak najszerszym zakresie, trudno sobie wyobrazić, że użytkownik będzie go ręcznie uzupełniał. Konieczne jest opracowanie metody automatycznego pozyskiwania informacji o aktywności użytkownika w różnych systemach oraz ich odwzorowywania w przyjętej ontologii. Jednym ze sposobów na pozyskiwanie takich informacji jest analiza dokumentów tekstowych czytanych przez użytkownika. Aby wyrazić je za pomocą ontologii, dla słów kluczowych wyekstrahowanych z czytanych przez użytkownika dokumentów konieczne jest przeprowadzenie procesu ujednoznaczniania pojęć.

2.1. Stan badań

Ujednoznacznianie pojęć (ang. *Word Sense Disambiguation*) jest jednym z obszarów badań nad przetwarzaniem języka naturalnego. Problem ujednoznaczniania pojęć został zdefiniowany już w późnych latach czterdziestych XX wieku jako jedno z zadań wchodzących w skład automatycznego tłumaczenia tekstów [Agirre i Edmonds 2006]. W kolejnym dziesięcioleciu przeprowadzono serię eksperymentów nad czynnikami, które pozwalają ludziom przypisywać odpowiednie znaczenia do słów w tekstach, które czytają. Czynnikiem takim jest na przykład uwzględnienie znaczeń sąsiednich słów lub gramatycznej struktury zdania, w którym dane słowo występuje. Kolejne dekady, wraz ze wzrostem mocy obliczeniowej komputerów, rozwojem metod sztucznej inteligencji oraz pojawieniem się wielu zasobów leksykalnych, przyniosły dalszy rozwój badań nad ujednoznacznianiem pojęć. Opracowano wówczas wiele podejść do rozwiązywania tego problemu, a niektóre z nich osiągnęły wysoki stopień poprawności ujednoznaczniania. Ciągle jednak ujednoznacznianie pojęć jest otwartym problemem informatyki i istnieje wiele kwestii wchodzących w jego zakres, które wciąż oczekują na rozwiązanie [Navigli 2009].

Współczesne podejścia do ujednoznaczniania pojęć można scharakteryzować, biorąc pod uwagę dwa kryteria, w zależności od tego, czy:

– nauka w takich podejściach przebiega w sposób nadzorowany, czy nienadzorowany,
– w metodach wykorzystywane są zewnętrzne zasoby wiedzy, czy też takich zasobów się nie wykorzystuje.

W pierwszym ze wspomnianych wymiarów metody rozróżnia się w zależności od tego, czy w procesie ich uczenia konieczne jest wykorzystanie ręcznie ujednoznacznionych przykładów – wówczas takie metody nazywamy nadzorowanymi (tzw. uczenie z nauczycielem). Przygotowanie takich przykładów dla metod nadzorowanych polega na ręcznym adnotowaniu słów w korpusie uczącym (słowom w takich korpusach ekspert nadaje odpowiednie znaczenia) [Broda i Mazur 2010]. Z kolei metody nienadzorowane (bez nauczyciela) nie wykorzystują przykładów uczących. Polegają one na identyfikacji częstych współwystąpień słów i przeprowadzanej na tej podstawie analizie skupień: poszczególne skupienia reprezentują tutaj odmienne znaczenia ujednoznacznianych słów. Dotychczasowe badania nad obiema grupami metod wskazują, że metody nadzorowane wykazują się większą dokładnością ujednoznaczniania [Navigli 2009], mają one jednak poważną wadę w postaci konieczności pracochłonnego przygotowania adnotowanych korpusów, wymaganych do przeprowadzenia nauki [Agirre i Soroa 2009].

Drugi wymiar rozróżnia metody z punktu widzenia wykorzystywania przez nie zasobów wiedzy do celów ujednoznaczniania. Takie zasoby to na przykład tezaurusy, przetwarzalne maszynowo słowniki, sieci semantyczne oraz ontologie. Przez długi czas brak takich zasobów i wysoki koszt ich przygotowania powodowały, że metody wykorzystujące zasoby wiedzy były wykorzystywane tylko do

tekstów dotyczących bardzo wąskich dziedzin [Ponzetto i Navigli 2010]. Dla tekstów w języku angielskim ta sytuacja zmieniła się w znacznym stopniu na początku lat dziewięćdziesiątych XX wieku, kiedy udostępniony został WordNet, leksykalna baza danych dla języka angielskiego, opracowana na Uniwersytecie Princeton. WordNet zawiera angielskie rzeczowniki, czasowniki, przymiotniki i przysłówki połączone w zbiory synonimów (tzw. synsetów), pomiędzy którymi ustalone są relacje, takie jak hiponimia, meronimia i antonimia [Miller 1995]. Od chwili udostępnienia, WordNet został wykorzystany w wielu projektach badawczych i może być uważany za *de facto* standard dla ujednoznaczniania pojęć w języku angielskim [Navigli 2009].

Z punktu widzenia artykułu szczególnie interesujące są metody, w których takie zasoby są wykorzystywane, ponieważ takie właśnie ujednoznacznianie pojęć może prowadzić do rozpoznania, które pojęcia w przyjętej ontologii reprezentują zainteresowania użytkownika. Popularnym przykładem metody ujednoznaczniania pojęć wykorzystującej zasoby wiedzy jest tzw. algorytm Leska [Banerjee i Pedersen 2010]. Określenie znaczeń dwóch słów występujących obok siebie w tekście odbywa się poprzez analizę ich definicji w wykorzystywanym słowniku, polegającą na policzeniu słów, które jednocześnie występują w definicjach obu tych wyrazów. Pojęcia, w których definicjach będzie najwięcej wspólnych wyrazów, przyjmowane są za właściwe znaczenia ujednoznacznianych słów. Wykorzystując takie podejście, osiągano trafność ujednoznaczniania rzędu 50–70%, w zależności od tego, jakie słowa podlegały ujednoznacznianiu [Navigli 2009].

Innym sposobem wykorzystania zasobów wiedzy w ujednoznacznianiu pojęć jest analiza relacji pomiędzy pojęciami w sieciach semantycznych, takich jak wspomniany wcześniej WordNet. Przykładem wnioskowania wykorzystującego takie relacje jest określenie, za pomocą pewnej metryki, semantycznego podobieństwa pomiędzy wszystkimi kombinacjami par znaczeń ujednoznacznianych słów występujących blisko siebie w tekście. Para pojęć, dla której takie podobieństwo jest największe, jest uznawana za właściwe znaczenia tych słów [Navigli 2009]. Obecnie coraz więcej uwagi poświęca się również metodom wykorzystującym w ujednoznacznianiu pojęć teorię grafów [Agirre i Soroa 2009]. Przykładem takiego podejścia jest praca G. Tsatsaronisa, I. Varlamis i K. Nørvåga [2010], w której wykorzystano algorytm PageRank w celu odkrycia i wykorzystania strukturalnych właściwości grafu, na którym oparty jest przyjęty zasób wiedzy. Słowa są tu ujednoznaczniane do pojęć na podstawie stopnia ich istotności ustalonego za pomocą takiego algorytmu.

2.2. Stan dziedziny dla języka polskiego

Omawiane w poprzedniej sekcji wyniki badań dotyczyły języka angielskiego. Należy zwrócić uwagę, że stan rozwoju ujednoznaczniania pojęć (oraz przetwarzania języka naturalnego w ogólności) różni się pomiędzy językami. Dla języka polskiego problem ujednoznaczniania pojęć stanowi znacznie większe wyzwanie niż

dla języka angielskiego, między innymi z powodu mniejszej dostępności zasobów lingwistycznych oraz charakterystyki samego języka (na przykład takich jego cech, jak bogata fleksja oraz w dużej mierze dowolny szyk wyrazów w zdaniu [Przepiórkowski 2007]).

Jeśli chodzi o zasoby leksykalne dla języka polskiego, w latach 2005–2008 na Politechnice Wrocławskiej prowadzono prace nad utworzeniem polskiego odpowiednika WordNetu, o nazwie SłowoSieć [Derwojedowa i in. 2008]. SłowoSieć jest zasobem mniejszym od WordNetu, jednak pomimo to stanowi poważny przełom w możliwości przeprowadzenia badań nad ujednoznacznianiem pojęć dla języka polskiego. Inne zasoby leksykalne istniejące dla języka polskiego to Narodowy Korpus Języka Polskiego (opracowany w ramach projektu finansowanego przez Ministerstwo Nauki i Szkolnictwa Wyższego w latach 2008–2010), w którym przeprowadzono ręczne ujednoznacznienie sensów słów dla podkorpusu składającego się z miliona segmentów [Przepiórkowski i Murzynowski 2009], polska DBPedia[3], analizator morfologiczny *Morfeusz*[4] oraz tager języka polskiego TaKIPI [Piasecki 2007], przeprowadzający ujednoznacznianie opisu morfosyntaktycznego wyrazów.

Obecne prace w zakresie ujednoznaczniania pojęć toczyły się bądź wciąż się toczą w dwóch ośrodkach: na Politechnice Wrocławskiej oraz w Instytucie Podstaw Informatyki PAN. W pierwszym z tych ośrodków przeprowadzono badania nad wybranymi metodami ujednoznaczniania opracowanymi dla języka angielskiego w celu ustalenia, jak będą działać dla języka polskiego. D. Bas, B. Broda i M. Piasecki [2008] zaprezentowali wyniki eksperymentów, w których ocenili trzy metody nadzorowanego ujednoznaczniania pojęć zastosowane dla tłumaczenia trzynastu polskich słów, reprezentujących różne problemy, specyficzne dla ujednoznaczniania pojęć w języku polskim. Rezultaty eksperymentów były zachęcające: w trzech testowanych metodach osiągnięto odpowiednio 80,23, 68,08 oraz 71,36% poprawności.

Kolejny artykuł dotyczący ujednoznaczniania pojęć dla języka polskiego, powstały w ramach badań na Politechnice Wrocławskiej, przedstawili B. Broda i W. Mazur [2010]. Poddano w nim analizie sześć metod nienadzorowanego ujednoznaczniania pojęć. Wyniki eksperymentów zaprezentowanych w artykule były interesujące, ponieważ wskazywały podobną poprawność wygenerowanych skupień do skupień powstałych w wyniku metod nadzorowanych.

Ciekawą pracę z zakresu ujednoznaczniania pojęć przedstawili R. Młodzki i A. Przepiórkowski [2011]. Autorzy zaprezentowali w nim tzw. *Word Sense Disambiguation Development Environment*, czyli środowisko do prac nad rozwojem ujednoznaczniania pojęć, będące narzędziem do projektowania i przeprowadzania eksperymentów z tego zakresu. Narzędzie to zostało przetestowane dla ujednoznaczniania

[3] Więcej na ten temat na stronie http://pl.dbpedia.org/ [dostęp: 18.01.2013].

[4] Więcej na ten temat na stronie http://sgjp.pl/morfeusz/ [dostęp: 18.01.2013].

języka polskiego dla korpusu i metod, które były użyte we wspomnianej wcześniej pracy D. Basa, B. Brody i M. Piaseckiego [2008]. Prototyp środowiska został udostępniony i jest dostępny w ramach licencji GNU General Public License.

Podsumowując, badania nad ujednoznacznianiem pojęć dla języka polskiego są na wczesnym etapie zaawansowania. Jednocześnie w ostatnich latach udostępnionych zostało kilka nowych zasobów i narzędzi pozwalających przypuszczać, że w najbliższych latach pojawią się kolejne propozycje rozwiązań.

3. Konstruowanie tożsamości użytkowników – analiza problemu

Zagadnienia omówione w poprzednich sekcjach dają szanse na rozwiązanie problemów dotyczących możliwości skonstruowania systemu globalnych modeli użytkowników oraz zasilania ich informacjami o użytkowniku. W tej sekcji artykułu spróbujemy zarysować kierunki, w jakich można by prowadzić dalsze badania w tym zakresie.

3.1. Globalny model użytkownika

Pomimo zaawansowanego stanu badań nad semantycznym modelowaniem użytkowników należy zwrócić uwagę na to, że rozwiązania pozwalające na ponowne wykorzystanie modelu użytkownika w wielu niezależnych systemach wciąż są w zasadzie nieobecne na rynku. Wynikać to może z kilku powodów. Jednym z nich jest prawdopodobnie brak zdefiniowania odpowiedniego modelu biznesowego dla rozwiązań tego typu, który zachęciłby autorów systemów do otwarcia się na wykorzystywanie i wspieranie budowy takich modeli użytkowników. Znajomość potrzeb użytkowników może być uważana za bardzo cenny zasób i dzielenie się tą wiedzą z konkurentami może być traktowane przez graczy na rynku jako sprzeczne z ich interesem. Jak już jednak wspomniano, możliwość ponownego wykorzystania modeli byłaby bardzo przydatna z perspektywy użytkowników. Wiele korzyści możliwość taka przyniosłaby również twórcom nowych systemów, którzy dzięki temu mogliby ominąć problem tzw. zimnego startu. Szerokie wykorzystanie globalnych modeli użytkowników wciąż jest jednak przyszłością.

Od kilku lat istnieją już rozwiązania służące do zarządzania tożsamością użytkowników w Internecie, rozumianą w kontekście mechanizmów umożliwiających potwierdzanie tożsamości danej osoby w różnego rodzaju systemach i podczas przeprowadzania transakcji internetowych. Takie systemy to między innymi OpenID, Identity Metasystem czy też WebID[5]. Na przykład, dzięki OpenID użytkownicy mogą mieć pojedynczy, globalny identyfikator z przypisanym do niego hasłem, za pomocą którego mogą się logować do wielu różnych serwisów, implementujących

[5] Więcej informacji na ich temat na stronach: http://openid.net/, http://docs.oasis-open.org/imi/identity/v1.0/identity.html, http://webid.info/ [dostęp: 17.01.2013].

moduły odpowiedzialne za współpracę z OpenID. Być może poszerzenie OpenID o zarządzanie semantycznym modelem użytkowników, który byłby konstruowany niezależnie od tego, z jakich portali dany użytkownik korzysta (np. za pośrednictwem wtyczki do przeglądarki analizującej teksty artykułów czytanych przez użytkownika), byłoby krokiem w stronę wypracowania odpowiedniego modelu biznesowego. Wówczas, w trakcie rejestracji użytkownika do określonego portalu za pomocą OpenID, możliwe byłoby przesłanie do tego portalu dodatkowych informacji o zainteresowaniach użytkowników. System OpenID ma wiele otwartych implementacji[6], które można wykorzystać w celu skonstruowania prototypu rozwiązania oraz przeprowadzenia eksperymentów.

Kolejne trudności wynikają z wymagania, aby globalny model użytkownika mógł być wykorzystywany w wielu różnych systemach, z których każdy może przejawiać odmienne rozumienie określonych domen. Jest to dobrze znany problem, istotny również z punktu widzenia zastosowania ontologii, które z założenia mają być wspólnym, przyjętym przez wiele stron sposobem patrzenia na pewien wycinek rzeczywistości. W związku z tym poprawne działanie systemu tożsamości użytkowników wymaga zaistnienia jednej z poniższych sytuacji:

– wszystkie systemy chcące korzystać z globalnych modeli użytkowników godzą się na korzystanie z określonej ontologii, za pomocą której modelują swą domenę i zainteresowania użytkownika,
– każdy system korzysta z własnej ontologii, ale dla potrzeb wykorzystania tożsamości muszą zostać utworzone mediatory, tłumaczące pojęcia z prywatnej ontologii na ontologię przyjętą w tożsamości.

W obu tych wypadkach autorzy systemów muszą się jednak zgodzić na korzystanie z określonej, wspólnej ontologii, dla której można zdefiniować następujące wymagania:

– ontologia musi być znana, dobrze udokumentowana, łatwo dostępna i musi funkcjonować na rynku od dłuższego czasu, tak aby twórcy systemów byli skłonni zgodzić się na jej używanie (lub utworzenie do niej mediatorów),
– musi pokrywać jak najszersze spektrum możliwych zainteresowań użytkowników, tak aby możliwe było uchwycenie jak najbardziej kompleksowego obrazu potrzeb informacyjnych użytkownika,
– w idealnym przypadku powinna być wielojęzykowa lub umożliwiać translację między pojęciami w różnych językach.

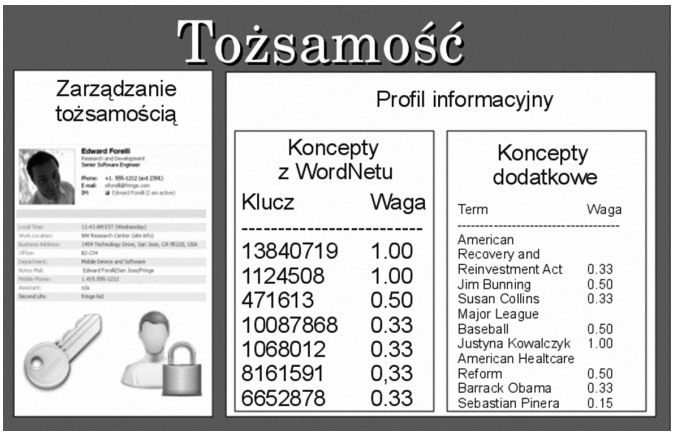
Przykładami ontologii spełniających większość lub wszystkie powyższe wymagania są: wspomniana w sekcji 1.2 WordNet wraz z polską wersją SłowoSieć, Cyc (zaproponowali ją D.B. Lenat i R.V. Guha [1991]) lub też coraz popularniejsza ostatnio DBPedia, będąca reprezentacją wiedzy zawartej w Wikipedii zgodną

[6] Przykładami takich implementacji dla języka Java są openid4java (http://code.google.com/p/openid4java/) lub WSO2 Identity Server (http://wso2.com/products/identity-server/ [dostęp: 17.01.2013].

ze standardem publikowania danych *Linked Data* (istnieje również jej polska wersja)[7]. Są one bardzo rozbudowane, a jednocześnie na tyle popularne, że mogłyby zostać uznane za standard reprezentacji obszarów zainteresowań użytkowników.

Kolejną kwestą konieczną do ustalenia jest sposób wykorzystania obranej ontologii w modelu użytkownika. Rozstrzygnięcia wymaga przede wszystkim kwestia tego, czy powinny być ustalane dodatkowe relacje między pojęciami modelującymi zainteresowania użytkownika. Relacje takie mogą być tworzone na przykład na podstawie ich współwystępowania, jak to miało miejsce między innymi w projekcie SiteIF [Stefani i Strappavara 1998]. Co prawda relacje między pojęciami istnieją już w ontologii, jednak wprowadzenie dodatkowych relacji może wzbogacić model o informację o tym, czy użytkownik interesuje się połączeniem kilku tematów. Na przykład, jeśli użytkownik interesuje się tematami „Programowanie w Javie", „Programowanie w PHP" oraz „Bazy danych (PostgreSQL)", jeśli istnieje relacja współwystępowania o wysokiej wadze między PHP a PostgreSQL, a nie ma takiej relacji między Javą a PostgreSQL, możliwe jest określenie, czy należy użytkownikowi rekomendować artykuły o wykorzystaniu bazy danych PostgreSQL w języku programowania Java. W badaniach nad globalnym modelem użytkownika należy przewidzieć możliwość modelowania dodatkowych relacji i określić eksperymentalnie, czy takie relacje rzeczywiście mają korzystny wpływ na wykorzystanie takich modeli.

W celu zaprezentowania możliwego rozwiązania w omawianym zakresie na rysunku 1 przedstawiono wizualizację struktury pierwszego prototypu



Rysunek 1. Wizualizacja struktury pierwszego prototypu wykorzystywanego w projekcie Ego

Źródło: Ekran projektu Ego

[7] Więcej informacji na ten temat na stronach: http://www.cyc.com/, http://dbpedia.org/About i http://linkeddata.org/ [dostęp: 17.01.2013].

tożsamości przygotowanego dla projektu Ego. Składa się on z następujących elementów:

– modułu odpowiadającego za zarządzanie tożsamością w rozumieniu analogicznym do OpenID, odpowiedzialnego za możliwości uwierzytelniania użytkowników na wielu różnych stronach,
– modelu użytkownika (zwanego tutaj profilem informacyjnym), składającego się z dwóch części:
  1) semantycznego modelu, składającego się z identyfikatorów pojęć (tutaj nazywanych konceptami) z WordNetu oraz przypisanych do nich wag,
  2) listy wyekstrahowanych z czytanych przez użytkownika słów kluczowych, których nie udało się przypisać (ujednoznacznić) do pojęć z ontologii.

3.2. Ujednoznacznianie pojęć dla potrzeb modelowania użytkowników – propozycja kierunku badań

Ujednoznacznianie pojęć z wykorzystaniem zasobów wiedzy jest jednym z najbardziej obiecujących sposobów na pozyskiwanie informacji o zainteresowaniach użytkowników sieci dla celów ich semantycznego modelowania. W przeszłości metoda taka była już wykorzystywana w badaniach zagranicznych, na przykład w omawianym w sekcji 2.1 projekcie SiteIF [Sosnovsky i Dicheva 2010]. Polega ona na tym, że jeśli w wielu artykułach czytanych przez danego użytkownika występują słowa ujednoznaczniane z określonym pojęciem w ontologii, to można uznać, że pojęcie to trafnie opisuje pewien obszar zainteresowań danego użytkownika.

W sekcji 2.1 jako podstawową przesłankę przypisania do słowa konkretnego zdarzenia wymieniono analizę kontekstu, w którym to słowo się znajduje (takim kontekstem może być na przykład kilka sąsiadujących z nim słów). Ciekawym rozszerzeniem tego pomysłu może być przyjęcie modelu użytkownika czytającego dany artykuł jako kolejnego elementu takiego kontekstu. Wówczas model użytkownika miałby wpływ na wyniki ujednoznaczniania. Aby zrozumieć przebieg i efekt takiego wnioskowania, można przeanalizować sytuację, w której należy określić znaczenie słowa „jaguar". Jeśli model osoby czytającej artykuł, w którym to słowo wielokrotnie występuje, wskazuje na wysokie zainteresowanie tej osoby motoryzacją, jako bardziej prawdopodobne znaczenie tego słowa należy obrać markę samochodów. Z drugiej jednak strony, wnioskowania takiego nie można przeprowadzać bezkrytycznie: możliwe jest, że wyjątkowo ta sama osoba czyta artykuł o jaguarze–drapieżniku. Wówczas takie wnioskowanie okazałoby się błędne. Informacja o zainteresowaniach użytkowników musi więc być wykorzystana w odpowiedni sposób – będzie ona jednym z czynników (oprócz na przykład wspomnianych sąsiadujących słów), który może pomóc ustalić odpowiednie znaczenie, muszą jednak również istnieć mechanizmy blokujące wyciąganie błędnych wniosków na tej podstawie.

Powyższy akapit wskazuje, że informacja o użytkownikach czytających dany artykuł może zostać wykorzystana w celu określenia właściwych znaczeń słów

pojawiających się w tekście. W odróżnieniu więc od klasycznego podejścia, gdzie ujednoznacznienie wpływa na modele użytkowników, w opisywanej sytuacji mamy zależność obustronną, gdzie również modele użytkowników wpływają na wyniki ujednoznaczniania pojęć.

Interesującym wyzwaniem może być dodatkowo wielokrotne ustalanie znaczeń słów w tym samym artykule wraz z tym, jak czytają go kolejne osoby. W przypadku gdy dany artykuł jest czytany najpierw przez użytkownika A, a następnie przez użytkownika B, a w modelach tych użytkowników słowo „jaguar" występuje w innych znaczeniach, wykorzystanie proponowanego rozumowania przyniosłoby odmienne wyniki ujednoznaczniania dla obu tych użytkowników. Pomysłem na rozwiązanie tego problemu jest zapisywanie dla każdego ujednoznacznianego wyrazu statystyk nadanych mu w przeszłości znaczeń. Wówczas, jeśli artykuł najpierw był czytany przez użytkowników A, B, C oraz D i każdy z nich był zainteresowany motoryzacją, to nawet jeśli kolejny czytelnik tego artykułu (użytkownik E) interesuje się drapieżnymi kotami, znaczenie słowa „jaguar" wciąż zostanie skierowane na markę samochodów. Dodatkowym wynikiem takiej analizy mogłoby być więc polepszenie dokładności semantycznego indeksowania dokumentów tekstowych.

4. Podsumowanie i dalsze kierunki badań

W artykule zaprezentowano przegląd badań z dwóch zakresów: semantycznego modelowania użytkowników i ujednoznaczniania pojęć. W obu tych dziedzinach prace toczą się już od dłuższego czasu. W artykule skupiono się na omówieniu najbardziej reprezentatywnych podejść, które można wykorzystać dla potrzeb modelowania wirtualnych tożsamości użytkowników. Dodatkowo, w sekcji 3 opisano wnioski, jakie wypływają z przeprowadzonej analizy wymagań dla wirtualnych tożsamości użytkowników oraz sposobów wykorzystania ujednoznaczniania pojęć dla ich konstruowania.

Dalsze kierunki badań będą się koncentrować wokół pogłębionej analizy możliwych do zastosowania zasobów wiedzy, mających być podstawą dla modeli użytkowników, oraz ujednoznaczniania pojęć, a także standardów, takich jak OpenID oraz OAuth, mających służyć za sposób komunikacji między serwerem tożsamości oraz systemami korzystającymi z tożsamości.

Bibliografia

Adomavicius, G., Tuzhilin, A., 2005, *Personalization Technologies: A Process-oriented Perspective*, Commun. ACM, vol. 48, s. 83–90, http://doi.acm.org/10.1145/1089107.1089109 [dostęp: 30.11.2012].

Agirre, E., Edmonds, P., 2006, *Word Sense Disambiguation: Algorithms and Applications*, Springer.

Agirre, E., Soroa, A., 2009, *Personalizing PageRank for Word Sense Disambiguation*, w: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA, s. 33–41, http://portal.acm.org/citation.cfm?id=1609067.1609070 [dostęp: 30.11.2012].

Anand, S.S., Kearney, P., Shapcott, M., 2007, *Generating Semantically Enriched User Profiles for Web Personalization*, ACM Trans. Internet Technol., vol. 7, http://doi.acm.org/10.1145/1278366.1278371 [dostęp: 30.11.2012].

Banerjee, S., Pedersen, T., 2010, *An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet*, Computational Linguistics and Intelligent Text Processing, s. 117–171.

Bas, D., Broda, B., Piasecki, M., 2008, *Towards Word Sense Disambiguation of Polish*, w: *International Multiconference on Computer Science and Information Technology, 2008. IMCSIT 2008*, IEEE, s. 73–78.

Bourdeau, J., Mizoguchi, R., 2000, *Using Ontological Engineering to Overcome Common AI-ED Problems*, International Journal of Artificial Intelligence in Education, vol. 11, s. 107–121.

Broda, B., Mazur, W., 2010, *Evaluation of Clustering Algorithms for Polish Word Sense Disambiguation*, w: *Proceedings of the International Multiconference on Computer Science and Information Technology*, IEEE, s. 25–32.

Brusilovsky, P., Millán, E., 2007, *User Models for Adaptive Hypermedia and Adaptive Educational Systems*, w: Brusilovsky P., Kobsa A., Nejdl W. (eds.), *The Adaptive Web*, vol. 4321, serie *Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, s. 3–53.

Cretton, F., La Calvé, A., 2008, *Generic Ontology Based User Model: GenOUM*, SMV technical report series, vol. 203.

Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., Broda, B., 2008, *Words, Concepts and Relations in the Construction of Polish WordNet*, w: *Proceedings of the Global WordNet Conference, Seged, Hungary*, Citeseer, s. 162–177.

Gruber, T., 1993, *A Translation Approach to Portable Ontology Specifications*, Knowledge Acquisition, vol. 5, no. 2, s. 199–220.

Heckmann, D., 2005, *Ubiquitous User Modeling*, vol. 297, IOS Press.

Kay, J., 1999, *Ontologies for Reusable and Scrutable Student Models*, AIED Workshop W2: Workshop on Ontologies for Intelligent Educational Systems, s. 72–77.

Lenat, D.B., Guha, R.V., 1991, *Ideas for Applying CYC*, MCC Technical Report ACT-CYC-407-91, December.

Magnini, B., Strapparava, C., 2001, *Using WordNet to Improve User Modelling in a Web Document Recommender System*, w: *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, vol. 31.

Miller, G.A., 1995, *WordNet: A Lexical Database for English*, Commun. ACM, vol. 38, s. 39–41, http://doi.acm.org/10.1145/219717.219748 [dostęp: 30.11.2012].

Młodzki, R., Przepiórkowski, A., 2011, *The WSD Development Environment*, Human Language Technology. Challenges for Computer Science and Linguistics, s. 224–233.

Nabeth, T., Gasson, M., 2005, *D2.3: Models*, FIDIS (Future of Identity in Information Society) Deliverable.

Nabeth, T., Hildebrandt, M., 2005, *D2.1: Inventory of Topics and Clusters*, FIDIS (Future of Identity in Information Society) Deliverable.

Navigli, R., 2009, *Word Sense Disambiguation: A Survey*, ACM Comput. Surv., vol. 41, s. 10:1–10:69, http://doi.acm.org/10.1145/1459352.1459355 [dostęp: 30.11.2012].

Niederée, C., Stewart, A., Mehta, B., Hemmje, M., 2004, *A Multi-dimensional, Unified User Model for Cross-system Personalization*, w: *AVI 2004*, Citeseer, s. 34.

Piasecki, M., 2007, *Polish Tagger TaKIPI: Rule Based Construction and Optimisation*, Task Quarterly, vol. 11, no. 1–2, s. 151–167.

Ponzetto, S.P., Navigli, R., 2010, *Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems*, w: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Association for Computational Linguistics, Stroudsburg, PA, s. 1522–1531, http://portal.acm.org/citation.cfm?id=1858681.1858835 [dostęp: 30.11.2012].

Przepiórkowski, A., Murzynowski, G., 2009, *Manual Annotation of the National Corpus of Polish with Anotatornia*, The proceedings of Practical Applications in Language and Computers (PALC-2009), Peter Lang, Frankfurt.

Przepiórkowski, A., 2007, *Slavonic Information Extraction and Partial Parsing*, w: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, ACL '07, Association for Computational Linguistics, Stroudsburg, PA, s. 1–10, http://portal.acm.org/citation.cfm?id=1567545.1567547 [dostęp: 30.11.2012].

Razmerita, L., Angehrn, A., Maedche, A., 2003, *Ontology-based User Modeling for Knowledge Management Systems*, User Modeling 2003, s. 148–148.

Salton, G., Wong, A. i Yang, C.S., 1975, *A Vector Space Model for Automatic Indexing*, Commun. ACM, vol. 18, s. 613–620, http://doi.acm.org/10.1145/361219.361220 [dostęp: 30.11.2012].

Sieg, A., Mobasher, B., Burke, R., 2010, *Improving the Effectiveness of Collaborative Recommendation with Ontology-based User Profiles*, w: *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, HetRec '10, ACM, New York, s. 39–46, http://doi.acm.org/10.1145/1869446.1869452 [dostęp: 30.11.2012].

Sosnovsky, S., Dicheva, D., 2010, *Ontological Technologies for User Modelling*, Int. J. Metadata Semant. Ontologies, vol. 5, s. 32–71, http://dx.doi.org/10.1504/IJMSO.2010.032649 [dostęp: 30.11.2012].

Stefani, A., Strappavara, C., 1998, *Personalizing Access to Web Sites: The SiteIF Project*, w: *Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia HYPERTEXT*, vol. 98, s. 20–24.

Tsatsaronis, G., Varlamis, I., Nørvåg, K., 2010, *An Experimental Study on Unsupervised Graph-based Word Sense Disambiguation*, w: Gelbukh, A. (ed.), *Computational Linguistics and Intelligent Text Processing*, vol. 6008, serie *Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, s. 184–198.

Zhang, H., Song, Y., Song, H., 2007, *Construction of Ontology-based User Model for Web Personalization*, User Modeling 2007, s. 67–76.

WORD SENSE DISAMBIGUATION OF THE POLISH LANGUAGE FOR THE NEEDS OF GENERATING DIGITAL IDENTITIES

Abstract: Nowadays, as more and more real world activities are transferred to the Web, the topic of digital identities, understood as representations of users' characteristics, become increasingly important. One of the purposes of research into digital identities is to use them as a means of informing the systems the user interacts with about his/her characteristics, needs and expectations, in order to, for example, enable a constant personalization process. However, a requirement that such identities are to represent a wide range of the user's characteristics in the form understandable to many different systems, raises many difficulties. An issue of how to collect such a broad set of information about users is also of particular significance. The aim of the article is to analyze two topics that can potentially be used while solving the above mentioned problems. These are ontology-based user modeling, which can be used for building reusable user models, and word sense disambiguation for words extracted from articles read by the users on different sites they visit. Based on the analysis of these topics, a proposal of further research directions is presented.