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

Elżbieta Bukowska, Monika Kaczmarek, Monika Walczak
Uniwersytet Ekonomiczny w Poznaniu, Wydział Informatyki i Gospodarki Elektronicznej, Katedra Informatyki Ekonomicznej
Autor do korespondencji: Elżbieta Bukowska, e.bukowska@kie.ue.poznan.pl

SEMANTYCZNIE WSPOMAGANE WYSZUKIWANIE EKSPERTÓW

Streszczenie: Obecnie o sukcesie organizacji często decyduje jej efektywność w obszarze zarządzania wiedzą. Potrzeba sprawnego wyszukiwania ekspertów wewnątrz i poza organizacją od dłuższego czasu jest inspiracją do podejmowania różnorodnych badań naukowych i inicjatyw ze strony przemysłu. Przykładem polskiego rozwiązania z tego zakresu jest system eXtraSpec. Aby możliwe było zrealizowanie efektywnego wyszukiwania ekspertów, konieczne było zaimplementowanie w systemie rozwiązań nie tylko pozwalających na pozyskiwanie i ekstrakcję informacji z różnych źródeł, lecz również umożliwiających operowanie na odpowiedniej reprezentacji danych, dzięki której możliwe jest wnioskowanie o charakterystykach opisujących osobę. Wykorzystane mechanizmy umożliwiają precyzyjną identyfikację potrzebnych danych, a równocześnie są wydajne i skalowalne. Artykuł prezentuje przebieg procesu wnioskowania w systemie eXtraSpec oraz przedstawia motywację i argumenty, które doprowadziły do takiej implementacji mechanizmu wnioskowania. Opisano w nim również ontologię stworzoną na potrzeby projektu oraz mechanizm indeksujący.

Słowa kluczowe: system wyszukiwania ekspertów, reprezentacja wiedzy, wnioskowanie.

Klasyfikacja JEL: D83, C52, M51.

Wstęp

Jako społeczeństwo informacyjne stajemy się coraz bardziej świadomi wartości rynkowej, jaką niesie ze sobą wiedza ekspercka. We współczesnym świecie nieustannie wzrasta zapotrzebowanie na ekspertów – specjalistów w określonej, zwykle wąskiej dziedzinie. Pracodawcy są gotowi ponieść znaczne koszty, aby rekrutować najlepszych specjalistów. Artykuł przedstawia semantycznie wspomagany mechanizm wyszukiwania ekspertów, będący częścią systemu eXtraSpec [Projekt eXtraSpec 2012], którego celem jest wspomaganie powyższych procesów. Podstawę do efektywnego wyszukiwania ekspertów stanowi odpowiedni opis eksperta, a więc

jego kompetencji i umiejętności, oraz opis historii zatrudnienia, w szczególności zajmowanych stanowisk i branż, w których pracował. Wyszukiwanie ekspertów w systemie polega na umożliwieniu przeszukiwania bazy zgromadzonych informacji w celu zidentyfikowania osób spełniających zdefiniowane w zapytaniu kryteria. Aby zapewnić odpowiednią jakość zwracanych rezultatów, do opisu profili ekspertów wykorzystano technologie semantyczne. W konsekwencji niezbędne stało się zdefiniowanie metod wnioskowania semantycznego. Wykorzystane w projekcie podejście zostało opisane w artykule. Szczegóły dotyczące poszczególnych jego etapów opisane zostały w kolejnych sekcjach.

1. Wyszukiwanie informacji i rola semantyki

Wyszukiwanie (ang. *information retrieval*) to proces, w którym pojedyncze zapytanie (ang. *query*) wykonywane jest w odniesieniu do pewnego zbioru dokumentów [van Rijsbergen 1995]. Przyjmuje się, że system wyszukiwawczy składa się z trzech komponentów:

- modułu odpowiedzialnego za gromadzenie danych (dokumentów) i za stworzenie ich reprezentacji w postaci indeksu dokumentu,
- interfejsu umożliwiającego zadanie zapytania; zapytanie przeważnie składa się z zestawu słów kluczowych i obrazuje aktualną potrzebę informacyjną użytkownika,
- mechanizmu dopasowującego zapytanie do zaindeksowanych dokumentów; w odpowiedzi na zadane zapytanie zwracana jest lista relewantnych dokumentów; komponent ten jest również odpowiedzialny za kolejność wyświetlanych wyników oraz ich graficzną prezentację.

Wszystkie wymienione powyżej elementy systemu wyszukiwawczego mają wpływ na jakość procesu wyszukiwania. Miarami jakości procesu wyszukiwania są: precyzja (ang. *precision*) i kompletność (ang. *recall*). Precyzję odpowiedzi definiuje się jako miarę będącą stosunkiem liczby dokumentów odnalezionych (tj. zaklasyfikowanych jako relewantne i faktycznie nimi będących) do całkowitej liczby dokumentów zakwalifikowanych do grupy relewantnych. Z kolei kompletność to stosunek liczby relewantnych dokumentów odnalezionych do całkowitej liczby dokumentów rzeczywiście relewantnych, które były analizowane przez system. W procesie wyszukiwania występują problemy uniemożliwiające osiągnięcie wysokich miar precyzji i kompletności wyników, takie jak:

- wykorzystywanie różnych słów kluczowych i różnego poziomu abstrakcji przez użytkowników przy formułowaniu zapytań dotyczących tego samego tematu,
- wykorzystanie różnorodnych słów i sformułowań w opisie danego zjawiska, tzn. w pozyskanych dokumentach i utworzonych na ich podstawie indeksach.

W przypadku gdy dopasowanie dokumentów do zapytania odbywa się za pomocą prostego mechanizmu, który sprawdza, czy w danym dokumencie występuje słowo kluczowe z zapytania, powyższe problemy skutkują:

- niską precyzją zwróconych wyników – duża liczba nierelewantnych dokumentów zwróconych przez system (występuje dane słowo, ale nie w danym kontekście),
- niską wartością miary kompletności – system nie identyfikuje relewantnych dokumentów znajdujących się w zbiorze dokumentów, gdyż opisane są innym zestawem słów kluczowych,
- dużą liczbą dokumentów zwróconych przez system (zwłaszcza w odpowiedzi na ogólne zapytanie), których przetworzenie przez użytkownika jest niemożliwe (np. ze względu na ograniczenie czasowe).

W celu rozwiązania powyższych kwestii, w procesie wyszukiwania stosuje się technologie semantyczne. Ich zastosowanie umożliwia systemom wyszukiwawczym zwracanie dokumentów, w których nie występują słowa zawarte w zapytaniu użytkownika, a które mimo to są relewantne (znaczeniowo) do zapytania.

Wprowadzenie technologii semantycznych do systemów wyszukiwawczych może przyjąć dwie formy:

- wykorzystania semantyki w celu analizy znaczenia indeksowanych dokumentów lub zadawanych zapytań,
- wykorzystania semantycznie opisanych zasobów w pracy systemu (tj. zasobach sieci semantycznej WWW).

Celem pierwszego podejścia jest dostarczenie zestawu relewantnych dokumentów dla zapytania poprzez porównanie znaczenia zapytania i treści dokumentu. W metodzie tej wykorzystana może być analiza lingwistyczna, identyfikacja najważniejszych pojęć bądź automatyczne lub półautomatyczne zawężenie lub poszerzenie kryteriów (ang. *query expansion*). Wyszukiwanie wykorzystujące analizę znaczeniową dokumentów eliminuje dokumenty zawierające terminy wyszukiwane, ale rozpoznane jako mające w danym kontekście odmienne znaczenie, oraz wprowadza do zbioru wyszukanych dokumentów te, które zawierają wyrazy o formie odmiennej od użytych w zapytaniu, ale semantycznie zbieżne. Analiza języka naturalnego, jak i proces rozszerzania bądź zawężania zapytania, wykorzystuje względnie stały słownik obejmujący zasady interpretacji poszczególnych terminów oraz wiedzę pozwalającą na wnioskowanie.

Drugi typ semantycznego wyszukiwania obejmuje systemy przeszukujące zawartość różnego typu plików RDF (Resource Description Framework) oraz ontologii, na przykład zapisanych w języku OWL (*Web Ontology Language*) i jest zgodny z ideą semantycznego Internetu [Berners-Lee, Hendler i Lassila 2001]. Semantyczne wyszukiwarki nie dokonują interpretacji semantycznej zawartości dokumentów z wykorzystaniem semantycznej analizy języka dokumentu, lecz bazują na już dostarczonym opisie dokumentu oraz odwołaniach do wybranych ontologii. Często korzystają również z mechanizmów wnioskujących (ang. *reasoners*) bazujących na zdefiniowanych ontologiach.

2. Systemy wyszukiwania ekspertów

Konieczność szybkiego odnalezienie w organizacji osoby bądź osób o odpowiednich umiejętnościach już od dłuższego czasu stanowi inspirację dla różnych inicjatyw mających na celu rozwój klasy wyszukiwarek wyspecjalizowanych w wyszukiwaniu ekspertów [Yimam 1996]. Z wyszukiwaniem ekspertów związane są między innymi dwa wyzwania [McDonald i Ackerman 2000]: identyfikacja kompetencji danej osoby prowadząca do udzielenia odpowiedzi na pytanie, „kto jest ekspertem w danej dziedzinie", oraz wybór ekspertów mający na celu zdefiniowanie, „co osoba X powinna wiedzieć". W naszych badaniach skupiamy się na pierwszym wyzwaniu, tj. identyfikacji osoby o odpowiednich kompetencjach.

Pierwsze systemy wyszukujące ekspertów wykorzystywały struktury bazodanowe zawierające syntaktyczny opis umiejętności ekspertów [Yimam-Seid i Kobsa 2003]. Systemy te jednak nie sprawdziły się w obliczu takich wyzwań, jak zapewnienie precyzyjnych wyników, gdy opis kompetencji jest ogólny, a zapytanie o ekspertyzę szczegółowe i ściśle określone [Kautz, Selman i Milewski 1996], czy też zagwarantowanie dokładności i wiarygodności informacji przechowywanych w statycznej bazie danych. Z tego powodu, w kolejnych systemach zaproponowano skupienie się na zautomatyzowanym pozyskiwaniu z określonych źródeł, na przykład z komunikacji e-mail, informacji dotyczących fachowej wiedzy poszczególnych osób [Campbell i in. 2003]. W dalszej kolejności zostały zaproponowane systemy, które indeksują i drążą opublikowane w intranecie dokumenty, traktując je jako źródła informacji na temat ekspertów [Hawking 2004; Metze, Bauckhage i Alpcan 2007]. Ponadto obecnie sieć WWW oferuje wiele możliwości znalezienia informacji na temat ekspertów, na przykład poprzez istniejące portale do zarządzania kontaktami lub portale społecznościowe, na których użytkownicy mogą szukać ekspertów czy potencjalnych pracowników, a z drugiej strony publikować swoje życiorysy, aby zostać znalezionym przez przyszłych pracodawców.

W pierwszych systemach wykorzystywano standardowe techniki wyszukiwania informacji [Ackerman, Wulf i Pipek 2002; Krulwich i Burkey 1996], w ramach których profil osoby zazwyczaj był reprezentowany jako wektor słów kluczowych, a rezultatem zapytania był wykaz osób spełniających zdefiniowane kryteria. Dopasowanie osoby do zapytania wykorzystywało prosty mechanizm sprawdzenia, czy dany wektor zawiera słowa kluczowe z zapytania, co doprowadziło do wspomnianych wcześniej problemów. Dlatego od kilku lat jest organizowana konferencja TREC (Text Retrieval Conference), której celem było zachęcenie naukowców do przeprowadzenia badań i eksperymentów w tej dziedzinie. Cykl konferencji w znacznym stopniu przyczynił się do rozwoju wiedzy w dziedzinie wyszukiwania ekspertów oraz zastosowania w niej licznych nowych technik i metod, na przykład metod probabilistycznych czy analizy języka naturalnego, w celu poprawienia jakości systemów wyszukujących [Balog i De Rijke 2006; Fang i Zhai 2007; Petkova i Croft 2006; Serdyukov i Hiemstra 2008].

Biorąc pod uwagę wciąż rosnącą popularność technologii semantycznych, nie zaskakuje to, że zostały one również wykorzystane w dziedzinie wyszukiwania ekspertów. Różne systemy wykorzystują ontologie do reprezentowania ofert pracy i życiorysów osób, np. w ramach projektu *Single Europeam Employment MarketPlace* [Gómez-Pérez, Ramírez i Villazón-Terrazas 2007] została opracowana ontologia wykorzystująca powszechnie stosowana standardy, takie jak ISO 4217, ISCO-88 (COM), ONET, do reprezentowania różnych elementów życiorysu danej osoby, np. kompetencji, aktywności zawodowej, wykształcenia. Kolejnym przykładem jest system ExpertFinder [Aleman-Meza i in. 2007], który dostarcza pojęć oraz wytycznych pozwalających adnotować strony internetowe, osoby, instytucje, wydarzenia, obszary kompetencji czy też wykształcenie. ExpertFinder korzysta z takich słowników, jak: FOAF, SIOC, vCard i Dublin Core. Ponadto w ramach różnych inicjatyw powstały liczne ontologie, taksonomie i klasyfikacje wspierające obszar zarządzania zasobami ludzkimi, umożliwiające na przykład opis stanowisk pracy SOC Federalnej Agencji Statystycznej Stanów Zjednoczonych czy też opis umiejętności, jak w projekcie KOWIEN [Dittmann 2003].

System omawiany w artykule należy do grona inicjatyw wykorzystujących semantykę w celu wyszukania eksperta w danej dziedzinie. System eXtraSpec pozyskuje informacje z zewnątrz organizacji w celu zbudowania profilu eksperta. System ten gromadzi informacje o dużej liczbie ekspertów, co z jednej strony oznacza większe pokrycie tematyczne i wzrost prawdopodobieństwa znalezienia odpowiednich osób do danego zapytania, jednakże powoduje również problemy związane z różnorodnością informacji, jak również dokładnością działania systemu. Wykorzystanie ontologii oraz wnioskowanie semantyczne mogą się przyczynić do normalizacji zebranych danych oraz zapewnienia odpowiedniego poziomu dokładności i kompletności mechanizmu. Ontologia opracowana dla potrzeb systemu eXtraSpec różni się od innych projektów tym, że: (1) nie ogranicza się tylko do stosunków hierarchicznych, (2) została opracowana dla języka polskiego i odnosi się do polskich norm; (3) została zbudowana z wykorzystaniem systemu SKOS [SKOS 2012].

W kolejnej sekcji przedstawione zostanie rozwiązanie związane z wnioskowaniem semantycznym dla projektu eXtraSpec oraz rozważane scenariusze.

3. Scenariusze wyszukiwania i rola semantyki

System eXtraSpec pozyskuje dane z wybranych źródeł, których zawartość jest zapisywana jako profile wyekstrahowane (PE), czyli pliki XML zgodne ze zdefiniowaną strukturą profilu. Słownictwo występujące w wyekstrahowanej treści jest następnie przetwarzane i normalizowane za pomocą zdefiniowanej ontologii. W efekcie z każdego wyekstrahowanego profilu powstaje odpowiadający mu strukturalnie profil znormalizowany (PN). Przyjmuje się, że jednemu rzeczywistemu ekspertowi może odpowiadać wiele znormalizowanych profili (np. pozyskanych z różnych źródeł lub

w różnych momentach). W celu ujednoznacznienia opisu danego eksperta konieczne jest połączenie informacji z wielu profili znormalizowanych i stworzenie profilu zagregowanego (PA). Zakładamy, że jednemu ekspertowi odpowiada jeden i tylko jeden profil zagregowany.

Aby umożliwić proces tworzenia profili, adnotowanie ich elementów oraz normalizowanie wartości, konieczne było stworzenie zestawu ontologii (słowników). Oprócz opisu, ontologie powinny umożliwić wnioskowanie na etapie wyszukiwania, a więc generowanie nowych faktów. Tematyka ontologii wykorzystywanych w projekcie została przedstawiona szczegółowo w sekcji 4.

Zakładając, że profil zagregowany został wzbogacony o semantyczne adnotacje oraz że istnieje użytkownik, który zadaje zapytania, należy rozważyć wykorzystanie semantyki na etapach:

1) tworzenia semantycznych indeksów pozyskanych profili zagregowanych, na podstawie których będzie realizowane wyszukiwanie,
2) zadawania zapytań – wzbogacenie zapytań o semantykę bądź przeprowadzenie analizy lingwistycznej lub rozszerzania zapytania (ang. *query expansion*),
3) tworzenia rankingu wyników.

W procesie tworzenia systemu wyszukiwawczego, oprócz dbałości o odpowiednią jakość zwróconych rezultatów (w rozumieniu precyzji i kompletności), bardzo istotne jest zapewnienie:

– odpowiedniej wydajności systemu (ang. *efficiency*) – rezultat wyszukiwania powinien być zwrócony w możliwie krótkim czasie,
– skalowalności systemu – obsługa rosnącej liczby zapytań i zaindeksowanych zasobów.

Biorąc pod uwagę powyższe kryteria, zidentyfikowane i rozważone zostały trzy możliwe scenariusze:

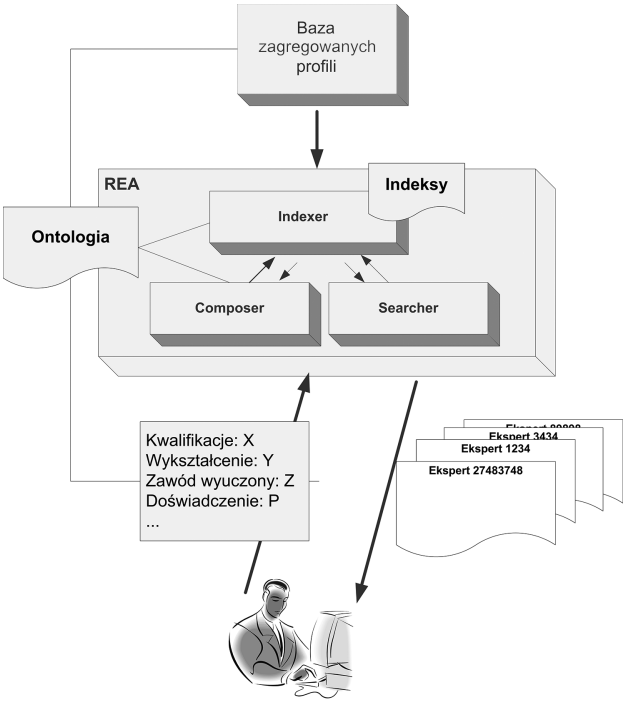
1. Wyrażenie zapytania z wykorzystaniem zdefiniowanej ontologii, a następnie wykonanie zapytania z wykorzystaniem mechanizmu wnioskującego. Podejście to wiąże się z koniecznością załadowania wszystkich ontologii do silnika wnioskującego oraz reprezentowania wszystkich profili jako instancji poszczególnych ontologii. Działanie na tak wyrażonych profilach i wnioskowanie na ich podstawie przez silnik wnioskujący jest procesem bardzo zasobochłonnym (czas dokonywania operacji wynosi nawet kilka minut, przy dużym zużyciu pamięci). Jest to rozwiązanie, które potencjalnie może się cechować wysoką precyzją i kompletnością, jednak niską wydajnością i skalowalnością systemu.
2. Rozszerzenie zapytania z wykorzystaniem ontologii (dodanie dodatkowych słów kluczowych do zapytania poprzez wykorzystanie ontologii, w celu zawężenia lub poszerzenia znaczenia oryginalnego zapytania). Podejście to zapewnia szybkie uzyskanie odpowiedzi i zwiększenie jej kompletności, nie jest jednak możliwe wzięcie pod uwagę dodatkowych relacji wyrażonych w ontologii. Zastosowanie tego podejścia może, ale nie musi skutkować zwiększeniem precyzji wyszukiwania.

3. Zbudowanie zapytania z wykorzystaniem zdefiniowanej ontologii jako kontrolowanego słownika, a następnie wykonanie zapytania na utworzonych wcześniej profilach (indeksach), wzbogaconych o dodatkowe informacje na etapie wstępnego wnioskowania. Zapytanie jest wykonywane na zbiorze profili za pomocą tradycyjnych mechanizmów wyszukiwania. Takie podejście nie wymaga korzystania z ontologii przy każdym zapytaniu, co pozwala na uniknięcie wad występujących w rozwiązaniu (1), przenosząc ciężar dokonania operacji na ontologiach na etap indeksowania.

Podsumowując, wyszukiwanie na podstawie słów kluczowych, chociaż obciążone niską precyzją i kompletnością zwróconych rezultatów ze względu na niejednoznaczności języka naturalnego, ma wielką przewagę nad innymi podejściami – mianowicie cechuje się wydajnością, skalowalnością oraz dojrzałością stosowanych mechanizmów. Zastosowanie trzeciego podejścia pozwoli wykorzystać mechanizmy z dziedziny wyszukiwania przy jednoczesnym zwiększeniu precyzji i kompletności zwróconych rezultatów poprzez:

– wprowadzenie etapu wnioskowania wstępnego (ang. *pre-reasoning*) w celu utworzenia wzbogaconych indeksów,
– minimalne wykorzystanie silnika wnioskującego w czasie wyszukiwania (tzn. tylko do kompozycji zespołów).

Realizowany scenariusz został przedstawiony na rysunku 1.



Rysunek 1. Architektura komponentu wnioskującego

Wyszukiwanie wraz z wnioskowaniem semantycznym jest realizowane przez komponent REA (ang. *REAsoning*), w ramach którego wyróżnić można przede wszystkim trzy istotne mechanizmy:

– mechanizm indeksujący (ang. *indexer*),
– mechanizm wyszukujący (ang. *searcher*),
– mechanizm kompozycji (ang. *composer*).

Mechanizmy te wykorzystują ontologie oraz odpowiedni silnik wnioskujący.

W kolejnych sekcjach artykułu przedstawiono struktury dla wnioskowania oraz architekturę systemu wnioskującego.

4. Struktury dla wnioskowania

Podstawą umożliwiającą działanie systemu oferującego semantycznie wspomagane wyszukiwanie ekspertów są ontologie. Odpowiednio przygotowana baza semantyczna zapewnia przedstawionemu rozwiązaniu skuteczniejsze i bardziej elastyczne możliwości wnioskowania niż w przypadku rozwiązań bazujących na prostym wyszukiwaniu ciągów znaków, bez powiązania ich znaczeń i zachodzących relacji.

Kluczowym etapem prac nad bazą ontologii było określenie formalizmu i modelu danych. Ostateczna decyzja dotyczyła ustalenia, w jakim stopniu tworzone rozwiązanie wymaga ekspresywności wykorzystywanego języka, a w jakim jego szybkiego przetwarzania. Obecnie nie istnieje formalizm, który spełnia obydwa te kryteria na wysokim poziomie. Przeprowadzona analiza wymagań wobec ekspresywności potencjalnego języka pozwoliła określić kluczowe i niezbędne do opisania relacje zachodzące pomiędzy pojęciami:

– hasSuperiorLevel: powiązania hierarchiczne,
– isEquivalent: relacja substytucji,
– isLocatedIn: różne zależności geograficzne,
– isLocatedInCity: zależności geograficzne,
– isLocatedInVoivodeship: zależności geograficzne,
– providesSkillDegree: powiązanie pomiędzy umiejętnością i certyfikatem,
– worksInLineOfBusiness: powiązanie pomiędzy organizacją i branżą,
– isPartOf: kompozycja elementów, np. umiejętność obsługa MS Word jest częścią składową umiejętności znajomość MS Office, (jednak znajomość MS Word nie implikuje znajomości całego pakietu MS Office).

Ponadto w ontologii wykorzystano relacje zdefiniowane przez SKOS: broader, hasTopConcept, inScheme, narrower, topConceptOf.

Dodatkowo ważnym aspektem przy wyborze formalizmu dla opisu ontologii w projekcie eXtraSpec jest powszechność języka. Zastosowanie standardu powszechnie znanego i stosowanego daje gwarancję odpowiednio dobrze rozwiniętego wsparcia programistycznego. Wymienione argumenty przyczyniły się do wyboru języka OWL.

Wymaganiem stawianym przed projektem eXtraSpec jest uniwersalność rozwiązania, umożliwienie szerokiego spektrum zastosowań. W kontekście modelowanej ontologii oznacza to opracowanie modelu prostego, a jednocześnie zawierającego wszystkie niezbędne elementy. Z tego powodu za model danych dla tworzonych na potrzeby projektu ontologiach przyjęto SKOS (ang. *Simple Knowledge Organisation System*) [SKOS 2012].

Wbrew niektórym wypowiedziom, SKOS nie jest językiem opisu ontologii, a modelem danych. Ontologie tworzone zgodnie z modelem SKOS mogą natomiast być zapisywane w dowolnym języku formalnym: od XML, przez RDF, aż do OWL. W specyfikacji SKOS opisany został jako ontologia w języku OWL. Natomiast konkretny model KOS (ang. *Knowledge Organisation System* – system organizacji wiedzy) jest reprezentowany jako instancja tej ontologii. Wybór formalizmu zapisu często determinuje wybór narzędzia wykorzystywanego do pracy z ontologią. Wszystkie ontologie wchodzące w skład stosu ontologii projektu eXtraSpec są tworzone z wykorzystaniem środowiska Protégé wzbogaconego o dodatek SKOSed, wspierający pracę ze SKOS-em.

Wiele systemów organizacji wiedzy, takich jak tezaurusy, taksonomie czy klasyfikacje, ma podobną strukturę i jest wykorzystywanych w podobnych aplikacjach. Model danych SKOS ujmuje większość tych podobieństw i precyzuje je, aby umożliwić wymianę danych i technologii pomiędzy różnymi aplikacjami. Zapewnia niskokosztową drogę migracji, pozwalającą na połączenie istniejących systemów organizacji wiedzy z semantycznym Internetem. Model SKOS zawiera wiele elementów, które można znaleźć w innych popularnych standardach organizacji wiedzy, takich jak BS czy też ISO 2788.

Zasadniczym elementem systemu eXtraSpec jest profil eksperta. Każdy ekspert opisywany jest zestawem informacji, na przykład są to imię i nazwisko, zdobyte wykształcenie, historia zatrudnienia, hobby, umiejętności, posiadane certyfikaty. Do przechowywania tych danych zaprojektowano odpowiednią strukturę – profil eksperta. W celu usprawnienia procesu wyszukiwania ekspertów w systemie eXtraSpec przeprowadzane jest wstępne wnioskowanie na profilach zagregowanych (PA), które poszerza te profile o nowe fakty, wykorzystując w tym celu utworzone ontologie (np. jeżeli ktoś ukończył Uniwersytet Ekonomiczny w Poznaniu, to znaczy, że ma wyższe wykształcenie) [Abramowicz i in. 2010]. Wnioskowanie takie jest możliwe dzięki utworzeniu swego rodzaju bazy wiedzy dziedzinowej, przechowującej informacje dotyczące tych atrybutów z profilu eksperta, które zostały uznane za istotne z perspektywy osoby poszukującej eksperta.

W projekcie przyjęto, że dla różnych zakresów informacji będą tworzone osobne ontologie. Otrzymany w ten sposób stos ontologii będzie bardziej uniwersalny i użyteczny również poza projektem eXtraSpec. Model SKOS zapewnia prostotę i łatwość tłumaczenia na inne formalizmy, jednocześnie wymusza wiele ustępstw, chociażby rezygnację z wielu udogodnień obecnych w języku OWL. Powstały stos ontologii, poza pojęciami, obejmuje również wspomniane wyżej relacje zachodzące

pomiędzy nimi. Tworzone ontologie musiałyby być zatem spójne z przyjętym modelem danych, przetwarzalne przez SKOS API, które jest wykorzystywane w projekcie, i w dalszym ciągu reprezentować wszystkie omówione wcześniej informacje. Aby sprostać tym wymaganiom, zaprojektowana baza wiedzy została zapisana jako jedna ontologia, która zawiera osiem schematów pojęć (ang. *ConceptShema*) reprezentujących poszczególne ontologizowane obszary. Do obszarów tych należą:

- organizacje edukacyjne (ang. *educational organization*) – uprawnione do nadawania określonych stopni i tytułów naukowych; ontologia zawiera listę polskich uczelni, wraz z informacją o ich kategorii (prywatna/publiczna) oraz typie (politechnika, uniwersytet, uczelnia ekonomiczna itp.),
- organizacje certyfikujące (ang. *certifying organization*) – uprawnione do nadawania różnego rodzaju certyfikatów; w opisie kompetencji użytkownika można wskazać posiadane przez niego certyfikaty wraz z instytucją, która dany certyfikat wydała,
- klient (ang. *client*), stanowisko (ang. *role*), organizacja zatrudniająca (ang. *employer*) – atrybuty wykorzystywane do opisu historii zatrudnienia eksperta; pojedynczy element historii zatrudnienia nazywany jest *relacją biznesową* (ang. *Business Relation*) i pozwala powiązać informacje o pracodawcy, branżę, stanowisko oraz obszary działania,
- zakres edukacji (ang. *scope of education*) – dziedzina edukacji (np. informatyka, inżynieria, nauki społeczne),
- temat edukacji (ang. *topic of education*) – nazwa specjalizacji (dla studiów wyższych) lub temat kursu (dla szkoleń),
- rezultat edukacji (ang. *result of education*) – uzyskany tytuł lub stopień,
- umiejętności (ang. *skill*) – klasyfikacja umiejętności, uwzględniająca ich wzajemne relacje, np. hierarchiczne,
- certyfikaty (ang. *name of certificate*) – możliwe do uzyskania certyfikaty,
- stopień opanowania umiejętności (ang. *degree of a skill*) – stopień znajomości danego zagadnienia.

W procesie normalizowania profilu wartości atrybutów z profilu wyekstrahowanego (PE) zostają połączone z pojęciami z ontologii. Istnieje możliwość, że mechanizm normalizujący nie znajdzie odpowiedniego pojęcia w ontologii. Taka wartość nie powinna być utracona, lecz zapisana w odpowiednim schemacie jako podpojęcie dla TMP (pojęć tymczasowych). Wartości tymczasowe będą weryfikowane przez eksperta i w razie potrzeby dodawane w odpowiednim miejscu do ontologii. W ten sposób możliwe jest systematyczne rozszerzanie ontologii o nowo pojawiające się trendy, produkty, firmy itp.

Źródła pojęć, które zostały wykorzystane do zbudowania poszczególnych ontologii, to:

- organizacje – gałąź ontologii opisująca organizacje edukacyjne została utworzona na podstawie listy polskich uczelni, opublikowanej przez Ministerstwo

Nauki i Szkolnictwa Wyższego [MNiSW 2012]; gałąź organizacji zatrudniających została przygotowana na podstawie różnych publikacji i spisów dostępnych w Internecie,
- stanowisko – klasyfikacja zawodów i specjalności opublikowana przez Ministerstwo Pracy i Polityki Społecznej [MPiPS 2012],
- zakres edukacji – ontologia jest kombinacją słownictwa stosowanego na różnych polskich portalach publikujących ogłoszenia o pracę,
- temat edukacji – lista specjalizacji, które są prowadzone na polskich wyższych uczelniach, zbudowana na podstawie materiałów opublikowanych przez Ministerstwo Nauki i Szkolnictwa Wyższego,
- rezultat edukacji – lista stopni zawodowych, tytułów i stopni naukowych możliwych do uzyskania w Polsce, zbudowana na podstawie odpowiednich rozporządzeń ministra nauki i szkolnictwa wyższego,
- certyfikaty – w prototypowym systemie ontologia zawiera listę certyfikatów językowych, możliwych do uzyskania przez obywateli Polski,
- umiejętności – listy klasyfikacji umiejętności stosowane na różnych polskich portalach publikujących ogłoszenia o pracę,
- stopień opanowania umiejętności – skala opanowania umiejętności stosowana na różnych polskich portalach publikujących ogłoszenia o pracę,
- miasto i województwa – lista polskich miast i województw,
- języki – lista języków, których znajomość może zostać potwierdzona certyfikatem w Polsce,
- branże – lista zbudowana na podstawie klasyfikacji wykorzystywanych przez różne polskie portale.

5. Semantycznie wspomagane wyszukiwanie ekspertów

Wyszukiwanie w bazie ekspertów wymaga implementacji dwóch niezależnych od siebie procesów:
1) tworzenia indeksów profili wraz z wnioskowaniem wstępnym,
2) zdefiniowania mechanizmu dopasowującego ekspertów do zadanych kryteriów wraz z odpowiednią konstrukcją zapytania.

Indeksowanie jest to proces tworzenia indeksu, czyli specjalizowanej bazy, zawierającej pewien wyciąg z indeksowanego dokumentu. Indeks powinien być zoptymalizowany pod kątem wyszukiwania, tj. zorganizowany tak, aby umożliwić bardzo szybkie wyszukiwanie na podstawie kryteriów zadanych przez użytkownika. W projekcie eXtraSpec tworzony jest on w przypadku pozyskania nowego profilu bądź zmiany w już istniejącym zbiorze profili. Proces tworzenia indeksu przebiega w następujący sposób: dany profil zagregowany (PA) jest analizowany, dzielony na odpowiednie sekcje, a następnie wzbogacony o dodatkowe informacje

z wykorzystaniem ontologii (ang. *pre-reasoning*). Zmiana indeksów konieczna jest przy każdej zmianie ontologii.

Drugi proces jest inicjowany przez użytkownika, który zadaje zapytania z wykorzystaniem graficznego interfejsu. Pracodawca, konstruując zapytanie, wskazuje na interesujące go kryteria oraz wartości, które zwracany profil powinien posiadać. Użytkownik, wybierając pożądane wartości poszczególnych cech z list i pól kombi, wskazuje na konkretne obiekty z ontologii. Następnie zapytanie jest zapisywane i przekazywane do komponentu wnioskującego, tj. REA. Jeśli zapytanie dotyczy komponowania zespołów, oryginalne zapytanie użytkownika jest rozbijane na kilka podzapytań. W wypadku wyszukiwania kryteria są bezpośrednio wykorzystane do konstrukcji zapytania. Zapytanie wykonywane jest na zbiorze zaindeksowanych profili oraz szeregowane zgodnie z semantycznym podobieństwem. W ostatnim kroku system zwraca listę identyfikatorów ekspertów spełniających kryteria zdefiniowane w zapytaniu.

Istnieje wiele silników wyszukiwania *open source* dostępnych na rynku. Według przeprowadzonych testów [Singh 2009], najlepszym narzędziem, biorąc pod uwagę optymalizację rozmiaru indeksu oraz jakość zwracanych wyników (precyzję i kompletność), jest Apache Lucene. Apache Lucene to otwarta biblioteka Javy, która pozwala na efektywne indeksowanie i przeszukiwanie informacji tekstowych, niezależnie od ich formatu.

Indeks jest najważniejszym elementem technologii Lucene. Zawiera on informacje o dokumentach Lucene i umożliwia sprawne wyszukiwanie tych, które pasują do zapytania. Podstawowe operacje API Lucene to dodawanie dokumentu do indeksu, usunięcie go i modyfikacja. Lucene wykorzystuje klasę *StandardAnalyzer* do badania tekstu (stemmer Portera) i wybierania słów do indeksowania. Istnieje również stemmer dla języka polskiego. W zakresie wyszukiwania, Lucene potrafi obsłużyć złożone wyrażenia zawierające operatory logiczne AND, OR, NOT oraz umożliwia m.in. maskowanie [*, ?] i wyszukiwanie fraz. Ze względu na elastyczność i możliwości rozbudowy, jak i dostosowania rozwiązań oferowanych przez Lucene do własnych potrzeb, właśnie to rozwiązanie zostało wykorzystane w projekcie do implementacji mechanizmów REA.

Pola w dokumencie w Lucene nie mogą być grupowane ani tworzyć hierarchii. W profilu zagregowanym (PA), na którym przeprowadzane jest wyszukiwanie, możliwe jest wyróżnienie hierarchicznej struktury, na którą składają się poszczególne elementy. Ponieważ niemożliwe jest odwzorowanie tej struktury w dokumencie Lucene, profile podczas indeksowania są rozbijane na oddzielne dokumenty zawierające:

- podstawowe dane personalne (np. imię, nazwisko, numer telefonu, adres),
- historię edukacji,
- posiadane certyfikaty,
- posiadane umiejętności,
- spis publikacji,

– historię zatrudnienia i aktywności zawodowych,
– listę organizacji, do których dana osoba należy,
– listę zainteresowań.

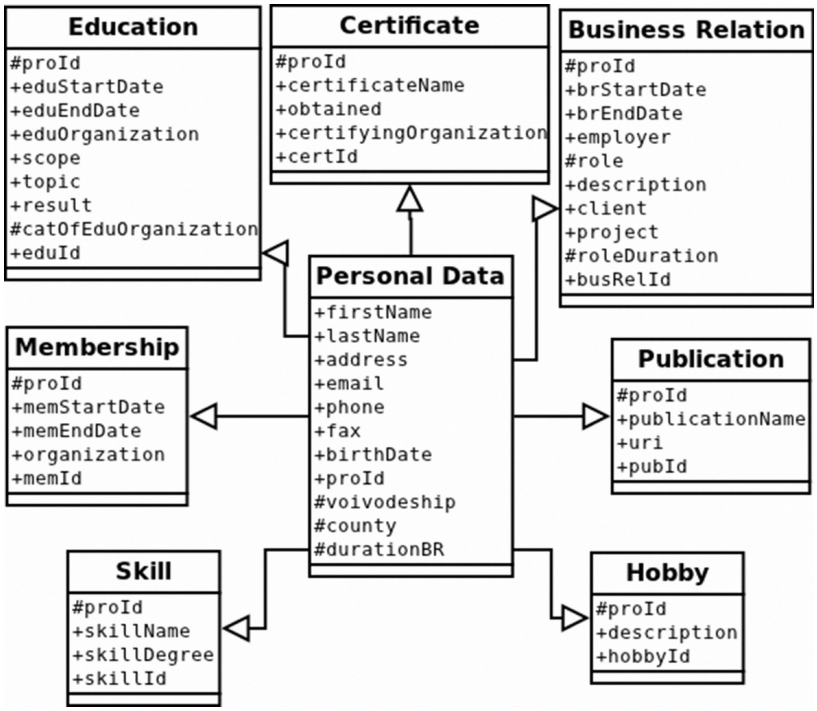
Wszystkie dokumenty mają dodatkowo pole zawierające identyfikator profilu, dzięki czemu wiadomo, którą osobę opisują. Dla każdego z wymienionych elementów tworzony jest oddzielny dokument, tzn. jednemu posiadanemu certyfikatowi odpowiada jeden dokument, podobnie dla każdego zajmowanego stanowiska w historii zatrudnienia tworzony jest oddzielny dokument itd.

W czasie indeksowania przeprowadzany jest również proces wstępnego wnioskowania (ang. *pre-reasoning*), w wyniku którego profil zostaje uzupełniony o dodatkowe fakty. Do tego celu wykorzystywany jest silnik wnioskujący z załadowanymi ontologiami omówionymi w poprzedniej sekcji. Standard URI elementu, dla którego odbywa się wnioskowanie, jest pobierany z profilu zagregowanego, a następnie przesyłany do silnika wnioskującego. Silnik ten zwraca wszystkie nadrzędne pojęcia dla danego elementu pobrane z odpowiedniej ontologii i przekazuje je do modułu indeksującego, zachowując właściwą hierarchię elementów. Pojęcia nadrzędne są zapisywane jako dodatkowe wartości danego pola. Zakładamy, że im bliżej w hierarchii położone jest dane pojęcie nadrzędne względem pojęcia podstawowego, tym większa powinna być jego waga. Ponieważ pojęcia nadrzędne dodawane są jako kolejne elementy tablicy przechowującej wartości danego pola, przyjmujemy, że im większy numer pozycji w tablicy, tym mniejsza waga pojęcia.

W przypadku niektórych elementów profilu pobrane pojęcia nadrzędne nie odpowiadają semantycznie elementom profilu. Wówczas do indeksowanego dokumentu dodawane są odpowiednie pola. Przykładem może być pole profilu zawierające adres, w którym na etapie wnioskowania można wyróżnić bardziej szczegółowe wartości: kod pocztowy, nazwę miasta zamieszkania czy nazwę ulicy. Na podstawie pobranego kodu pocztowego możliwe jest określenie powiatu i województwa, co pozwoli na przeprowadzanie wyszukiwania przy wykorzystaniu kryteriów geograficznych (bardziej ogólnych niż nazwa miasta). Ponieważ jednak profil eksperta nie zawiera wydzielonych pól przechowujących nazwę powiatu czy województwa, są one dodawane na etapie indeksowania do dokumentu zawierającego podstawowe dane personalne.

W dokumentach odpowiadających poszczególnym elementom edukacji dodano pole *catOfEduOrganization*. Dla każdej instytucji edukacyjnej z ontologii pobierane są również jej pojęcia nadrzędne, a w indeksie przechowywana jest cała pobrana hierarchia. Pozwala to na przykład na wyszukanie osób, które ukończyły informatykę na politechnice (przy czym nie interesuje nas, jaka konkretnie była to uczelnia).

W dokumentach zawierających posiadane umiejętności rozszerzono zawartość pola *skillName* w taki sposób, że są w nim przechowywane również wszystkie nadrzędne pojęcia umiejętności z ontologii. Podobna sytuacja występuje w dokumencie zawierającym poszczególne elementy historii zatrudnienia, gdzie pole *role* zostało poszerzone o hierarchię pojęć nadrzędnych dla zajmowanego stanowiska.



Rysunek 2. Schemat dokumentów Lucene podlegających indeksowaniu

Schemat dokumentów podlegających indeksowaniu i składających się na nie pól przedstawiony jest na rysunku 2.

Oprócz wzbogacenia tworzonego indeksu o informacje wywnioskowane z wykorzystaniem relacji hierarchicznych ('*is a*'), w przyszłości do indeksów dodawana będzie informacja wynikająca także z dodatkowych relacji zamodelowanych w ontologii.

Podsumowanie

System eXtraSpec opisany w artykule wspiera analizę dokumentów wewnętrznych przedsiębiorstwa oraz wybranych źródeł internetowych w celu wyszukiwania osób odznaczających się określonymi kompetencjami lub posiadających wiedzę w danej dziedzinie.

W systemie tradycyjne metody wyszukiwania zostały wzbogacone ontologiami. Zastosowanie technologii semantycznych pozwala na wyszukanie dokumentów relewantnych do zapytania nawet wówczas, gdy słowa kluczowe wykorzystane w zapytaniu nie odpowiadają dokładnie słowom kluczowym w dokumencie. Semantyka została wprowadzona zarówno na etapie budowania zapytań

(tworzonych na podstawie ontologii), jak również podczas opisywania zasobów (zastosowanie wnioskowania wstępnego pozwala uniknąć wnioskowania na ontologii podczas każdego zapytania, co znacznie zwiększa efektywność działania systemu), a także wykorzystywana jest podczas tworzenia rankingów wyników wyszukiwania.

Dane wyekstrahowane z dedykowanych źródeł są zapisywane w postaci dokumentu XML jako profil wyekstrahowany (PE). Następnie poszczególne elementy PE są normalizowane przez komponent NOR, który wykorzystuje leksykony i ontologie w celu stworzenia profilu znormalizowanego (PN). Profile znormalizowane są łączone w profile zagregowane (PA), tak aby jednej osobie odpowiadał jeden profil, zawierający informację o zmianach profilu w czasie.

Analizując możliwe do wykorzystania struktury dla wnioskowania, ostatecznie jako system indeksujący i wyszukiwawczy wybrano bibliotekę Apache Lucene. Indeksowanie (wraz z wnioskowaniem wstępnym) oraz opis semantyczny profili jest wykonywany z wykorzystaniem zbioru ontologii opisanych za pomocą języka OWL oraz zgodnych z modelem danych SKOS.

System został utworzony w celu przetwarzania dokumentów w języku polskim. Również ontologie są zbudowane z pojęć w języku polskim, jednak same mechanizmy oraz relacje są uniwersalne i nic nie stoi na przeszkodzie, aby zastosować je do analizy dokumentów w innych językach.

Bibliografia

Abramowicz, W., Kaczmarek, T., Stolarski, P., Wecel, K., Wieloch, K., 2010, *Architektura systemu wyszukiwania ekspertów eXtraSpec*, w: Gołuchowski, J., Frączkiewicz-Wronka, A. (red.), *Technologie wiedzy w zarządzaniu publicznym*, Wydawnictwo Akademii Ekonomicznej w Katowicach, Katowice s. 295–313.

Ackerman, M., Wulf, V., Pipek, V., 2002, *Sharing Expertise: Beyond Knowledge Management*, MIT Press.

Aleman-Meza, B., Bojars, U., Boley, H., Breslin, J.G., Mochol, M., Nixon, L.J., Polleres, A., Zhdanova, A.V., 2007, *Combining RDF Vocabularies for Expert Finding*, w: *Proceedings of the 4th European conference on The Semantic Web: Research and Applications*, Springer-Verlag, Innsbruck, Austria, s. 235–250, http://portal.acm.org/citation.cfm?id=1419662.1419685\&coll=portal\&dl=ACM [dostęp: 22.03.2012].

Balog, K.L.A., De Rijke, M., 2006, *Formal Models for Expert Finding in Enterprise Corpora*, w: *Proceedings of the ACM SIGIR*, 3–7 June 2007, Innsbruck, Austria, s. 43–50.

Berners-Lee, T., Hendler, J., Lassila, O., 2001, *The Semantic Web*, http:www.scientificamerican.comarticle.cfm?id=the-semantic-web [dostęp: 20.05.2009].

Campbell, C.S., Maglio, P.P., Cozzi, A., Dom, B., 2003, *Expertise Identification Using Email Communications*, w: *CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*, ACM Press, s. 528–321.

Dittmann, L., 2003, *Towards Ontology-based Skills Management – KOWIEN*, Projektbericht zum Verbundprojekt KOWIEN, Institute for Production and Industrial Management, University Duisburg-Essen.

Fang, H., Zhai, C., 2007, *Probabilistic Models for Expert Finding*, w: *ECIR'07 Proceedings of the 29th European Conference on IR Research*, Springer-Verlag, Berlin–Heidelberg, s. 418–430.

Gómez-Pérez, A., Ramírez, J., Villazón-Terrazas, B., 2007, *An Ontology for Modelling Human Resources Management Based on Standards.*, w: Apolloni, B., Howlett, R.J., Jain, L.C. (eds.), *KES* (*1*), vol. 4692, seria *Lecture Notes in Computer Science*, Springer, Vietri sul Mare, Italia.

Hawking, D., 2004, *Challenges in Enterprise Search*, w: *Proceedings of the 15th Australasian Database Conference – Volume 27*, ADC '04, Australian Computer Society, Darlinghurst, Australia, s. 15–24, http://portal.acm.org/citation.cfm?id=1012294.1012297 [dostęp: 22.03.2012].

Kautz, H., Selman, B., Milewski, A., 1996, *Agent Amplified Communication*, w: *Proceedings of the Thirteenth National Conference on Artificial Intelligence* (*AAAI-96*), AAAI Press, Portland, Oregon, s. 3–9.

Krulwich, B., Burkey, C., 1996, *ContactFinder Agent: Answering Bulletin Board Questions with Referrals*, w: *Proceedings of the National Conference on Artificial Intelligence*, AAAI Press, Portland, Oregon, s. 10–15.

McDonald, D.W., Ackerman, M.S., 2000, *Expertise Recommender: A Flexible Recommendation System and Architecture*, w: *CSCW'00: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, ACM Press, Philadelphia, s. 231–240.

Metze, F., Bauckhage, C., Alpcan, T., 2007, *The "Spree" Expert Finding System*, w: *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*, IEEE Computer Society, s. 551–558.

MNiSW, 2012, *Uczelnie*, http://www.nauka.gov.pl/szkolnictwo-wyzsze/system-szkolnictwa-wyzszego/uczelnie/ [dostęp: 22.03.2012].

MPiPS, 2012, http://www.praca.gov.pl/pages/klasyfikacja_zawodow2.php [dostęp: 22.03.2012].

Petkova, D., Croft, W., 2006, *Hierarchical Language Models for Expert Finding in Enterprise Corpora*, w: *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, IEEE Computer Society, Washington D.C., s. 599–608.

Projekt eXtraSpec, http://extraspec.kie.ue.poznan.pl/ [dostęp: 30.12.2012].

Serdyukov, P., Hiemstra, D., 2008, *Modeling Documents as Mixtures of Persons for Expert Finding*, w: *Proceedings of the ECIR*, Springer Verlag, Glasgow, s. 309–320.

Singh, V., 2009, *A Comparison of Open Source Search Engines*, http://zooie.wordpress.com/2009/07/06/ [dostęp: 30.12.2012].

SKOS 2012, http://www.w3.org/TR/swbp-skos-core-spec [dostęp: 22.03.2012].

Rijsbergen, C.J. van, 1995, *Information Retrieval and Information Reasoning*, *Computer Science Today*, vol. 100, s. 549–559.

Yimam, D., 1996, *Expert Finding Systems for Organizations: Domain Analysis and the Demoir Approach*, w: *ECSCW 999 Workshop: Beyond KNowledge Management: Managing Expertise*, ACM Press, New York, s. 276–283.

Yimam-Seid, D., Kobsa, A., 2003, *Expert Finding Systems for Organizations: Problem and Domain Analysis and the Demoir Approach*, Journal of Organizational Computing and Electronic Commerce, vol. 13, no. 1, s. 1–24.

SEMANTICALLY-ENABLED RETRIEVAL OF EXPERTS

Abstract: Nowadays, efficient utilization of knowledge has become the key to success of an organization. The need to find experts within or outside an organization has for a long time been an inspiration for various types of research as well as industrial initiatives. An example of expert finding systems is the Polish initiative called eXtraSpec. In order to realize its tasks, the eXtraSpec system needs not only to be able to acquire and extract information from various sources, but also requires an appropriate representation of information supporting the reasoning as regards a person's characteristics. The considered mechanism should allow for a precise identification of the required data, and simultaneously be efficient and scalable. This paper presents the reasoning scenario used within the eXtraSpec project and discusses the underlying motivation which led to the development of the semantically enabled pre-reasoning mechanism. The developed ontology as well as implementation details of the indexing mechanism are also discussed.