Semantic Job Portal
SemJob

(vedoucí: Martin Nečaský, Ph.D.; necasky@ksi.mff.cuni.cz)

Počet řešitelů: 5

Motivace:
Na Internetu existuje celá řada portálů pro vyhledávání v nabídkách práce, např. sprace.cz, jobs.cz, atd. Ty však nabízejí pouze základní vyhledávací funkce: (1) částečná shoda na základě vyplněného vyhledávacího dotazníku a (2) fulltextové vyhledávání klíčových slov. Důsledkem jsou nízké hodnoty ukazatelů „precision“ a „recall“, které jsou v oblasti dokumentografických informačních systémů používány k měření kvality. Zcela zásadním důvodem tohoto stavu je fakt, že současné portály žádným způsobem nepracují se sémantikou. Nerozumějí pracovním nabídkám a životopisům uloženým v jejich databázích. Stejně tak nerozumějí zadávaným vyhledávacím dotazům. Bez znalosti sémantiky je navíc obtížné vytvářet další nadstavbové aplikace, jako např. aplikace pro sledování trendů v dané oblasti pracovního trhu. Doplnění sémantiky ve strojově čitelném formátu by tedy mohlo výrazně vylepšit výsledky vyhledávání na současných pracovních portálech.

Základní myšlenka doplnění sémantiky do existujícího vyhledávacího systému je z teoretického pohledu poměrně jednoduchá. Stačí doplnit ontologii, která sémantiku popisuje, a dokumenty anotovat podle této ontologie. Současné sémantické dotazovací jazyky (např. SPARQL) již pak nabízejí řadu zajímavých nástrojů pro sémantické zpracování takových dokumentů. Problémem je však technická realizovatelnost takového rozšíření.

Cíl projektu:
Cílem projektu je vytvoření funkčního experimentálního portálu pro sémantické vyhledávání v pracovních nabídkách. Pracovat budeme v úzké spolupráci se společností dlouhodobě působící na pracovním trhu. Pro účely projektu se zaměříme pouze na jednu oblast pracovního trhu, konkrétně na informatiku.

První částí projektu bude vytvoření poloautomatického nástroje pro anotaci pracovních inzerátů a životopisů oproti ontologii vybraného segmentu trhu. Dále bude implementováno úložiště takto anotovaných dokumentů. Logicky bude úložiště založeno na datovém modelu RDF a bude nabízet kromě ukládání také modul pro dotazování v některém ze sémantických dotazovacích jazyků (SPARQL nebo podobný). Další částí projektu bude modul pro sémantické fulltextové vyhledávání – zadaný fulltextový dotaz bude mapován na ontologii a převeden tak na sémantický dotaz nad RDF daty v databázi. Poslední částí projektu bude vytvoření sady ukázkových aplikací nad anotovanými dokumenty, např. aplikace pro sledování trendů ve vývoji nejžádanějších znalostí uchazečů o informatické povolání.

Vstupy do projektu budou následující (budou vytvořeny ve spolupráci s partnerskou společností a nebudou hodnocenou součástí projektu):

1. ontologie popisující vybranou oblast pracovního trhu (pravděpodobně informatiku)
2. korpus reálných pracovních inzerátů a životopisů
3. sada požadavků na funkcionalitu, kterou by doplnění sémantiky mělo umožnit

Vymezení částí projektu:

1. Anotační modul bude založen na existujících algoritmech počítačové lingvistiky, jako např. lemmatizace atd. Samotná anotace bude probíhat na základě pravidel. Budou existovat dva typy pravidel: (a) ručně vytvořená a do systému dodaná pravidla a (b) pravidla, která se systém naučí na základě vstupního korpusu. Modul tedy bude také implementovat metody strojového učení pravidel. (2 členové týmu)
2. Modul pro ukládání anotovaných dokumentů bude založen na existujícím databázovém systému. Bude se pravděpodobně jednat o kombinaci existujících volně dostupných úložišť pro XML data a RDF data. Pokud bude zjištěno, že žádné dostupné úložiště není plně vyhovující (např. výkonnostně), přistoupíme k jejich úpravě, např. doplnění indexačních metod pro zvýšení efektivity dotazování. (1-2 členové týmu)
3. Sémantické full-textové vyhledávání bude založeno na lingvistických metodách a pravidlech podobně jako (1). (1 člen týmu)
4. Ukázkové aplikace budou tvořit webové rozhraní k vytvořenému systému. Budou pracovat s datovým modelem RDF. Zde se může ukázat, že současná programovací API (např. Jena) nejsou dostatečná a bude je potřeba optimalizovat. (1 člen týmu)

Předpokládaný průběh práce:

1. Analýza existujících implementací a přístupů
2. Podrobná specifikace konkrétních funkcí systému, architektury a rozhraní mezi jednotlivými moduly
3. Implementace projektu
4. Testy na reálných schématech, ladění
5. Dokumentace (programátorská, uživatelská, instalační)