Sémantický web pro novináře
PressSemWeb

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

Noviny a časopisy typu Respekt, Týden, a mnoho dalších dnes ve svých článcích rozkrývají celou řadu osob, skupin a vztahů, které měly zůstat skryty před veřejností. V takové síti je velmi složité se orientovat ať už pro samotné autory článků – novináře, tak pro jejich čtenáře.

Problémem je, že sít je zachycena ve veřejné podobě pouze ve formě samotných článků. Zajímavou myšlenkou tedy je zachycení sítě vztahů do informačního systému, který by umožňoval svým uživatelům tuto síť libovolně rozšiřovat o nové osoby, skupiny, vztahy atd. To by umožnilo přirozeným způsobem integrovat znalosti novinářů v jednom otevřeném systému. Navíc by měl systém potenciál nabídnout daleko pokročilejší způsoby vyhledávání informací, než je jen „klasické“ vyhledávání výskytů klíčových slov na webových stránkách. To by umožňovalo uživatelům nejen se v síti lépe orientovat, ale také objevovat nové skryté vztahy daleko efektivněji. Bylo zjištěno, že o takový systém je zájem jak mezi novináři, tak i čtenáři.

Cílem tohoto studentského softwarového projektu je experimentální systém, který by takovou síť umožnil zachytit, libovolně rozšiřovat a efektivně prohledávat. Obecně lze již nyní říci, že datovým modelem bude (multi-)graf. Nelze ale předem zafixovat všechny konkrétní typy uzlů a hran. Jistě zde budou předem identifikovatelné typy uzlů (např. *osoba*, *skupina*) a hran (např. *podplatil, dohodil, podvedl*). Je však nutné, aby byl systém flexibilní pro přidávání nových typů samotnými uživateli. Kromě možnosti rozšiřování je také potřeba poskytnout pokročilejší možnosti vyhledávání v podobě nějakého uživatelsky přívětivého dotazovacího jazyka nad grafy s různými typy hran a uzlů. Netriviální bude návrh vhodného uživatelského rozhraní. To bude muset být dvojúrovňové. Na obecné úrovni to bude přívětivé rozhraní pro procházení obecného grafu. Na konkrétní úrovni to budou specializovaná rozhraní pro speciální typy uzlů a hran. To klade netriviální požadavky na rozšiřitelnost uživatelských rozhraní „za běhu“.

Navržené technické řešení nemůže být pouze přívětivé pro uživatele. Další nutností je efektivní uložení zachycených informací a efektivní vyhodnocování dotazů.

Systém musí být přirozeným rozšířením současného webu, musí v sobě integrovat existující webové stránky a provazovat je mezi sebou či s novými typy uzlů pomocí nových typů hran (které jsou tedy vlastně rozšířením „klasických“ webových linků). Důležitá bude znalost sémantiky jednotlivých uzlů a hran. Sémantika musí být explicitně vyznačena přímo v datech. Dnes existuje rozsáhlý výzkum v oblasti tzv. Sémantického webu (http://www.scientificamerican.com/article.cfm?id=the-semantic-web). Ten na teoretické úrovni řeší spoustu technických problémů takové sítě. Výsledky výzkumu se proto v práci pokusíme využít (např. technologie RDF, OWL, sémantické dotazovací jazyky SPARQL a pod., ...).

V projektu nebudeme řešit právní stránku věci. Výstupem bude experimentální systém, který se pokusí odpovědět na některé technické problémy a bude moci sloužit jako prototyp pro případná řešení použitelná v praxi. Pokusíme se také navázat spolupráci přímo s novináři.