Vyhledávání v elektronických informačních zdrojích

Vlastimil Krejčíř, ÚVT MU

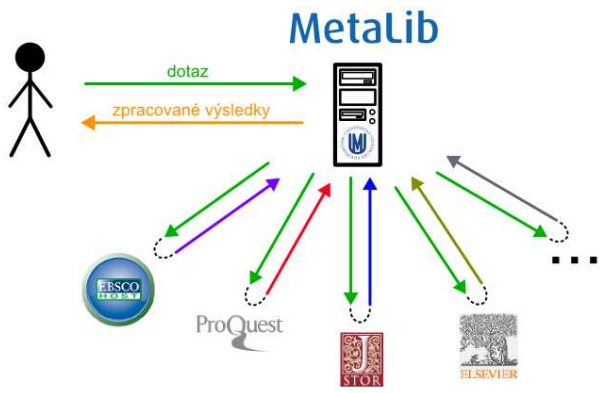
Jednou z možných metod práce s elektronickými informačními zdroji je vyhledávání na základě zadaného dotazu. Ten je možné vložit do formuláře přímo v daném informačním zdroji přes jeho nativní webové stránky. Tento způsob ale není vhodný v případě, kdy uživatel neví, ve kterém zdroji se hledané informace nacházejí. A prohledávat na zadaný dotaz každý zdroj zvlášť je velmi pracné a časově náročné. Na řadu tak přichází nástroje, které dokáží prohledat více informačních zdrojů najednou (tzv. paralelní vyhledávání).

V současné době se užívají dva základní přístupy k řešení problému paralelního prohledávání elektronických informačních zdrojů. První způsob je postaven na principu prostředníka vyřizujícího dotazy za uživatele – bývá také označován jako federativní vyhledávání a je v praxi na akademických institucích (především ve světě) poměrně rozšířen. Druhý vyznává filozofii á la Google, tedy budování velkého indexu. Tento přístup je v oblasti elektronických zdrojů relativně nový. V následujícím textu se pokusíme podívat na oba přístupy podrobněji a ukázat si jejich výhody i slabé stránky.

1 Federativní vyhledávání (prostředník)

Federativní vyhledávání staví mezi uživatele a elektronické zdroje třetí stranu – jakéhosi prostředníka, který za uživatele udělá „špinavou práci“ – dotazy vyřídí za něj. V praxi je tímto prostředníkem nějaký software, ke kterému mohou uživatelé přistupovat přes webový prohlížeč. Na Masarykově univerzitě je uživatelům takový software k dispozici – jmenuje se Metalib[1])

Jak celý proces vyhledávání v praxi probíhá? Uživatel vloží do formuláře na webu k vyhledání nějaký dotaz. Metalib jej přebírá a postupně přeposílá na jednotlivé informační zdroje. Na nich dochází k vyhodnocení a vyhledání výsledků – během této doby Metalib pouze čeká, než mu



Obrázek 1: Princip federativního vyhledávání (systém Metalib)

zdroje vrátí výsledky. Ihned poté, co Metalib začne výsledky přijímat, začíná s jejich zpracováním. Nejdříve je převede do jednotného formátu, sloučí je a setřídí (např. podle relevance, data, autorů apod.). Následně takto zpracované výsledky přehledně zobrazí uživateli, který tak obdrží homogenní a přehledný seznam. Celý proces je schematicky ukázán na obrázku 1.

Na první pohled se tento postup vyhledávání zdá být ideálním řešením. Podívejme se však podrobněji, co všechno musí Metalib udělat a co z toho pro uživatele plyne.

Výhodou federativního vyhledávání je především to, že se hledá v aktuálních datech – ve výsledcích je přesně to, co by uživatel nalezl při ručním prohledávání jednotlivých zdrojů. Prostředník hraje pouze pasivní roli stran toho, co je samotným obsahem prohledávaných zdrojů. Změní-li některý zdroj poskytovaný obsah, pak se z pohledu prostředníka nic neděje – uživatel stále dostává přesně to, co je právě ve zdroji k dispozici.

Velkou nevýhodou z pohledu uživatele může být rychlost; celý proces využití prostředníka trvá poměrně dlouhou dobu – část času se spotřebuje čekáním na výsledky (doba je u každého zdroje jiná a obvykle se čeká na zdroj nejpomalejší, což mohou být i desítky sekund). Část je spotřebována na zpracování výsledků – a ta také klade velké nároky na prostředníka: čím většího počtu zdrojů se dotazuje, tím víc výsledků

[1] `http://metalib.muni.cz/`

musí zpracovat. V praxi se ukazuje, že je vhodné jedním dotazem prohledávat jen omezený počet zdrojů (kolem 12; v Metalibu-MU jsou již předchystány skupiny zdrojů dle oborového zaměření jednotlivých fakult, uživatel si také může vytvářet skupiny vlastní, více viz [1]). Při větším počtu už dochází k přetížení prostředníka (serveru) a doba zpracování dotazu se velmi prodlužuje. Pro zrychlení celého procesu zpracovává Metalib pouze prvních 30 výsledků z každého zdroje – ty jsou obvykle nejrelevantnější. Další výsledky je však možno na přání uživatele zpracovat a zobrazit také.

Samotný Metalib musí především jednotlivé informační zdroje přesně znát a přistupovat k nim individuálně – vědět, jakým způsobem se jich dotazovat a v jakém formátu mu budou tyto zdroje vracet výsledky. Každý zdroj navíc nemusí být „přátelský“ a poskytovat výsledky ve strojově snadno zpracovatelném tvaru. Proto nastupuje člověk (administrátor), který příslušná napojení na jednotlivé zdroje nastaví a Metalib nakonfiguruje. Čas od času pak musí sledovat, jestli daná napojení fungují a nedošlo-li na straně některého zdroje ke změnám. Kromě větší pracnosti to může ovlivnit i uživatele – ne na každý zdroj je možné snadno udělat napojení, v případě změny na straně zdroje může docházet k občasným výpadkům.

Shrňme si výhody a nevýhody federativního vyhledávání:

+ aktuálnost a přesnost výsledků
- doba odezvy
- náročnost na výkon prostředníka
- nutná aktivní správa

2 Velký index á la Google

Použití velkého indexu je běžné na poli internetových vyhledávačů. Internetové vyhledávače jsou tu již léta a technologie, kterou používají, je tedy poměrně dobře ověřena. Okamžitě vyvstává otázka, proč totéž nepoužít i pro oblast elektronických informačních zdrojů. Jak si ukážeme dále, situace není v případě elektronických zdrojů tak jednoduchá.

Jak vlastně klasické vyhledávače typu Google (nebo Seznam.cz apod.) pracují? Jednou z jejich základních činností je sběr dat. Vyhledávač neustále brouzdá Internetem a stahuje si obsah všech stránek, na které narazí. Stažený obsah si pak indexuje a ukládá do vlastní databáze (která nabývá poměrně značných rozměrů). Většinu stránek Google navštěvuje opakovaně, sleduje změny, ke kterým na nich došlo, a tyto si průběžně aktualizuje ve své databázi. Když uživatel zadá dotaz, vyhledávač pracuje pouze s vlastní vybudovanou a dobře indexovanou databází, což výrazně urychluje dobu reakce, která je v podstatě okamžitá.

Použití stejné technologie u bází elektronických zdrojů (zde je známa pod názvem *vyhledávací služby* – discovery services) by odstranilo nevýhody federativního vyhledávání. Prostředník by již nebyl pouze pasivním prvkem, ale stal by se inteligentním partnerem, který shromažďuje veškerá data z elektronických informačních zdrojů a buduje si z nich vlastní databázi. Zkrátí tak dobu odezvy na dotaz na minimum, podstatně také sníží nároky na výkon při dotazu – prohledávání lokální databáze je méně náročné než dotazování se třetích stran, navíc její budování probíhá předem a nezávisle na dotazech uživatelů. Ve srovnání s federativním vyhledáváním tedy ušetříme čas, po který se čeká na odpověď od zdrojů, a také čas i výkon nutný k zpracování a sloučení odpovědí (to vše je již předem připraveno ve vlastní databázi).

Prostředníkovo postavení v takovém systému musí být nutně silnější. Bude vyžadován lepší hardware, především rychlá datová úložiště pro samotné uložení indexovaného obsahu, a také patřičně výkonný server, který bude zvládat techniky práce s velkým indexem. Tyto požadavky je dnes možné splnit.

Prostředník musí být ale silnější i na poli „politickém“. Shromažďuje data, která nejsou běžně a zdarma na Internetu dostupná. V tom se liší od běžných vyhledávačů typu Google. Ty sbírají data, která jsou volně dostupná komukoli. Náš prostředník je ve složitějším postavení a musí si vyjednat souhlas poskytovatelů elektronických zdrojů k přístupu a následnému shromažďování a indexování dat. V praxi se toto ukazuje jako největší překážka – jednotliví producenti často

nejsou ochotni poskytovat svá data přímo třetím stranám. Mnozí z nich jsou sami agregátory a poskytují vyhledávání nad větším množstvím zdrojů – tito si pak často chtějí ponechat exkluzivitu pro dané zdroje a v rámci obchodního boje data svým konkurentům přirozeně neposkytnou. Často se jedná zejména o zdroje velké a pro úplné vyhledávání nezbytné. V současné době spolu na trhu soupeří několik velkých hráčů, což uživatelům situaci komplikuje.

Menší nevýhodou je nižší aktuálnost vyhledávání – lokální index nemůže okamžitě kopírovat veškeré změny obsahu, které u jednotlivých poskytovatelů nastávají. Aktualizace lokálního indexu může probíhat jen s určitou periodicitou, v mezidobí index zastarává. Pokud například poskytovatel zdroje vymění větší část obsahu, pak index na straně prostředníka může po nějakou dobu podávat neplatné informace.

Na Masarykově univerzitě máme centrální index také k dispozici, a to v rámci systému Metalib. Uživatel si může vybrat vyhledávání ve skupině „Primo Central“ a jeho dotaz je vyřízen okamžitě právě díky indexu v lokální databázi. Index je budován firmou ExLibris, která je autorem Metalibu. Zároveň je jedním z hráčů na poli vyhledávacích služeb; jeho přímými konkurenty jsou EBSCO (vyhledávací služba Ebsco Disvovery service), Serials Solution (vyhledávací služba Summon) a některé další.

Shrňme si výhody a nevýhody vyhledávání přes velký index:

+ rychlost odezvy
+ menší náročnost na výkon prostředníka
- aktuálnost výsledků
- nutná „politická“ dohoda s producenty zdrojů

3 Závěr

Jak jsme si ukázali, oba přístupy k vyhledávání mají své kladné i záporné stránky. Budoucnost bude nejspíše patřit velkým indexům – nabízí výborný výkon a technické překážky jsou u nich již vyřešeny. Zásadní bude to, jak dopadne konkurenční boj mezi poskytovateli exkluzivního obsahu.

V současnosti připravujeme knihovní projekt do evropských strukturálních fondů, jehož součástí bude i pořízení velkeho indexu pro MU.

Literatura

[1] V. Krejčíř. Nástroje pro práci s elektronickými informačními zdroji MU. Zpravodaj ÚVT MU. ISSN 1212-0901, 2010, roč. XX, č. 3, s. 3-7. http://www.ics.muni.cz/bulletin/articles/634.html □