Nové možnosti vyhledávání vědeckých lékařských informací
(New Possibilities for Scientific Medical Information Retrieval)

Vendula Papíková [1], Richard Papík [2]

[1] Centrum biomedicínské informatiky, Ústav informatiky AV ČR, v.v.i., Praha

[2] Ústav informačních studií a knihovnictví, Univerzita Karlova, Praha

papikova@euromise.cz, richard.papik@ff.cuni.cz

INFORUM 2007: 13. konference o profesionálních informačních zdrojích

Praha, 22. - 24. 5. 2007

Souhrn:

Internet jakožto jedna z cest k odborným informacím stále nabývá na významu. Hnutí za otevřený přístup k vědeckým informacím je jednou z příčin takového vývoje a významně se podílí na zvyšování využitelnosti webu v odborných a vědecko-výzkumných kruzích. Nicméně informační přetížení, heterogenita webového prostředí a specifické informační potřeby jednotlivých uživatelů znesnadňují efektivní vyhledávání dostatečně relevantních informací. Takto posilovaná potřeba po "stále chytřejších" vyhledávacích nástrojích a navíc současně přicházející druhá generace webových služeb označovaná jako Web 2.0 mění tvář internetu jako celku a odráží se i v sektoru pokrývajícím odborné a vědecko-technické informace.

Toto sdělení pojednává o nových, volně dostupných i placených vyhledávacích systémech přístupných prostřednictvím Internetu a o webových službách souborně označovaných jako Web 2.0 se zaměřením na vyhledávání vědeckých informací v biomedicínských oborech. Prezentovány jsou vyhledávací nástroje využívající shlukování ("clustering") a vizualizaci vyhledaných dokumentů, hluboké indexování („deep indexing“), fenomén tzv. „social bookmarking“, folksonomie a nové aplikace vyvíjené pomocí aplikačních rozhraní vyhledávačů, tzv. APIs ("application programming interfaces"), jako jsou alternativní uživatelská

rozhraní do biomedicínckých databází nebo tzv. „mushups“. Uvedeny jsou rovněž konkrétní příklady výše zmíněných vyhledávacích systémů, rozhraní i služeb.

Práce byla podpořena projektem 1M06014 MŠMT ČR a projektem IAA701010606 GA AV ČR.

Klíčová slova: *vědecké lékařské informace, internet, online vyhledávání, web 2.0*

Abstract:

An importance of the Internet as a source of professional information is growing. The Open Access movement to scientific information is one of the reasons of this trend and significantly improves usability of the web in professional and academic quarters. However information overload, web heterogeneity and specific information needs of users make an efficiency of information retrieval difficult. Need of „ever smarter“ searching tools and the second generation of web services denoted as Web 2.0 changes the face of Internet as complex and is reflected also in area covering professional, scientific and technical information.

This poster deals with new, free accessible, as well as licensed information retrieval systems and with web services collective denoted as Web 2.0 in scope of the scholarly information in biomedical sciences. They are presented search engines playing upon clustering and visualization of found documents, deep indexing, social bookmarking, folksonomy and new applications developed via APIs (application programming interfaces), as are alternative user´s interfaces of biomedical databases or so called mushups. The concrete examples of retrieval systems, interfaces and services are demonstrated.

This work was supported by a project 1M06014 MŠMT ČR and by a project IAA701010606 GA AV ČR.

Key words: *scientific medical information, internet, online information searching, web 2.0*

Úvod

Současně s přibývajícím množstvím vědeckých publikací na jedné straně a s nárůstem specifických informačních potřeb uživatelů na straně druhé vzniká poptávka po nových sofistikovaných nástrojích pro jejich efektivní prohledávání. Medicína a související obory přírodních věd patří mezi ty oblasti lidského poznávání, kde je exponenciální růst nových informací zvláště akcentován. Zákonitě z toho plynoucí „atomizace“ medicínských oborů vede k úzké specializaci a dokonce až k „superspecializaci“ lékařů (př. invazívní kardiologové, horní či dolní endoskopisté). Vznikají nové obory a nová paradigmata (př. nutrigenomika a nutrigenetika, medicína založená na důkazech). Současně sílí potřeba interdisciplinárního přístupu. Informační potřeby jednotlivých specialistů se tak stále více vymezují a diferencují. Mění se také informační chování uživatelů jako odraz Web 2.0. Tato skutečnost je hybnou silou pro vznik nových informačních zdrojů a také pro vznik nástrojů pro jejich efektivní prohledávání.

Inovativní vyhledávací technologie a přístupy

Inovativní vyhledávací technologie a přístupy lze rozdělit do následujících kategorií:

- **„Vertikalizace“** obecných (horizontálních) internetových vyhledávačů:
  Poté, co Google uvedl svůj úspěšný projekt **Google Scholar** (http://scholar.google.com), pustil se do sféry vyhledávání vědecké literatury také Live Search a svůj nástroj nazval **Live Search Academic** (http://academic.live.com).

- **Shlukování** (cluster search):
  Příkladem z této kategorie je vyhledávač **ClusterMed** (http://demos.vivisimo.com/clustermed), který využívá technologii Vivísimo k seskupování výsledků vyhledaných v databázi MEDLINE/PubMed. Patří tedy současně také mezi alternativní vyhledávací rozhraní (viz dále) pro tento klasický biomedicínský informační zdroj.

- **Vizualizace** (visual search):
  Nástroj pro seskupování a následnou vizualizaci vyhledaných, obsahově souvisejících dokumentů má například vyhledávač **Grokker**, který je implementován do systému databází **EBSCO** (http://search.ebscohost.com). Vizualizace vztahů mezi souvisejícími články je také jednou z funkcí rozhraní HubMed (http://www.hubmed.org), které umožňuje alternativní přístup k databázi MEDLINE/PubMed.

- **Podrobná indexace** (deep indexing):
  Například databáze **CSA Illustrata** (http://www.csa.com, http://info.csa.com/csaillustrata) indexuje a prohledává tabulky, vzorce, mapy, grafy, diagramy, schémata a další vyobrazení, která jsou součástí vědeckých článků.

- **Metavyhledávání** (federated search):
  Nástroje z této kategorie umožňují prohledávání dvou a více informačních zdrojů prostřednictvím jediného rozhraní.

- **Automatická analýza textu** („meta-grafy“, text / literature mining):
  Mezi zástupce této kategorie patří v medicínském kontextu vyhledávač databáze Evidence Matters (http://www.evidencematters.com), který analyzuje data z publikací klinických studií a znázorňuje je ve formě přehledných tabulek a „meta-grafů“. Umožňuje tak rychle určit terapeutické možnosti u konkrétních pacientů se stanovenou diagnózou.
  Jiným příkladem je rozhraní PubReMiner (http://bioinfo.amc.uva.nl/human-genetics/pubreminer), jehož zdrojovou databází je MEDLINE/PubMed. PubReMiner analyzuje tituly a abstrakta článků odpovídající zadanému dotazu a generuje z nich frekvenční tabulky. Dle četnosti v sestupném pořadí jsou v nich uvedeny roky, kdy byly články publikovány, autoři, jejich pracoviště, názvy časopisů, v nichž práce vyšly, termíny MeSH a názvy substancí. Díky nim je možné dotaz snadno a velmi efektivně redefinovat. Z tabulky je lze také vyčíst, jakým tématem se zabývá určitý autor, na jakém pracovišti se věnují

určitému tématu nebo do kterého časopisu je nejvhodnější poslat práci na to či ono téma.

- **Inkorporace oborových taxonomií, ontologií a řízených slovníků** do vyhledávacích rozhraní:
  Na tomto místě jsou vhodným příkladem rozhraní **GoPubMed** (http://www.gopubmed.org) a **MeSHPubMed** (http://www.meshpubmed.org). Jená se o alternativní rozhraní do databáze MEDLINE/PubMed. Genová ontologie (resp. MeSH) v nich slouží jako systematická struktura pro více než 17 miliónů článků obsažených v MEDLINE/PubMed. Technologie uvedených rozhraní třídí abstrakta vyhledaná ve zdrojové databázi pomocí genové ontologie, resp. MeSH.

- Služby a aplikace **Web 2.0**, jejichž příklady budou uvedeny dále.

Webové služby a aplikace druhé generace (Web 2.0)

Jak již bylo zmíněno, kromě vzniku specifických informačních potřeb úzce specializovaných uživatelů dochází i k postupné změně jejich informačního chování díky vlivu prostředí Web 2.0. Webové služby a aplikace druhé generace kladou důraz na otevřenost (dostupnost) a podporují komunikaci, spolupráci a sdílení informací. Pro Web 2.0 jsou charakteristické především tyto vlastnosti:

1. **Zveřejňování rozhraní pro programování aplikací** (APIs application programming interfaces), která umožňují vývoj **nových uživatelských rozhraní** a „smíšených aplikací“ (**mushups**).

   Uživatelská rozhraní tak **mohou být přizpůsobována** specifickým potřebám konkrétních skupin uživatelů (**inkorporace funkcí, integrace zdrojů, zvyšování uživatelské přívětivosti**). Například pro databázi MEDLINE/PubMed t.č. existuje několik desítek alternativních rozhraní.
   Jedním z nich je rozhraní PubMed Interact (http://pmi.nlm.nih.gov/interact), které díky svým intuitivním a interaktivním prvkům (systém posuvných jezdců)

usnadňuje zadávání vyhledávacích limitů do té míry, že není potřeba znát pravidla dotazovacího jazyka, ani ztrácet čas zdlouhavým zaškrtáváním vyhledávacích limitů ve formulářích známých z tradičního rozhraní PubMed (www.pubmed.gov). Řada dalších příkladů alternativních rozhraní do MEDLINE/PubMed je zmíněna na jiných místech tohoto článku.

Mushups jsou webové stránky nebo aplikace kombinující obsah ze dvou nebo více externích online zdrojů do jednoho integrovaného nástroje. Zdrojem obsahu pro mushup bývají obvykle jiné webové stránky např. cestou API nebo RSS/Atom.

K medicínským aplikacím z této kategorie patří například EpiSPIDER (http://www.epispider.org). Tento nástroj integruje několik elektronických zdrojů z oblasti monitorování infekčních nemocí a akutních otrav. Vizualizuje mj. hlášení ze systému ProMED-mail pomocí Google Maps.

2. Systémy pro vytváření **společných záložek** v rámci komunity uživatelů (social bookmarks) a **folksonomií** (tj.„taxonomií“ vytvářených komunitou uživatelů), často vizualizovaných v podobě tzv. **oblaku tagů** (tag cloud, buzzcloud), které umožňují využití „kolektivní inteligence“ (collective intelligence, wisdom of crowds).

   Výhody: Jsou pružné, nesvázané pevně danou strukturou, intuitivní a uživateli blízké. Jsou užitečným „doporučujícím“ nástrojem (odhalují aktuální a „horká témata“) a jsou také predikujícím nástrojem (indikují trendy).
   Úskalí: Vznikají spontánně, neorganizovaně a nestrukturovaně, proto se na ně nelze spoléhat při systematickém průzkumu daného informačního zdroje nebo tématu.

   Oblak tagů (tag cloud, buzzcloud) vizualizuje zájem o témata vyjádřená názvy tagů, jeho základem však mohou být také vyhledávaná klíčová slova, příslušné termíny MeSH apod. Nejvíce zastoupená slova jsou v něm zvýrazněna největším písmem, nejméně hledané termíny jsou znázorněny nejmenším písmem.

   Příkladem z této kategorie je webová služba **CiteULike** (http://www.citeulike.org), pomocí které lze ukládat a organizovat odborné

články. Ty pak mohou být prohledávány, hodnoceny a komentovány celou komunitou uživatelů tohoto systému.

3. **Zaměření na uživatele**: Je možné vytvářet vyhledávače přizpůsobitelné individuálním zájmům jednotlivých uživatelů bez nutnosti znalosti programování (**customized search**). Zajímavým nástrojem je také uživateli posílené vyhledávání (**user powered search**).
   Například technologie **Swicki** zohledňuje znalosti a preference online komunit a dodává tak váhu a vertikální specifitu vyhledaným výsledkům. Charakteristika přizpůsobeného vyhledávání (customized search) je naplněna tím, že systém přednostně prohledává stránky zadané autorem vyhledávače. Charakteristika uživateli posíleného vyhledávání (user powered search) je splněna tím, že autor i členové dané online komunity mohou vyhledané stránky hodnotit. S ohledem na jejich hodnocení jsou modifikovány výsledky příštího vyhledávání (**učící se vyhledávač**). Technologie současně vizualizuje „hot searches“ (frekvenci vyhledávaných termínů korigovanou hodnocením výsledků hledání danou online komunitou) ve formě tzv. "**Buzzcloud®**“. Příkladem medicínského vyhledávače tohoto druhu je Addison´s Disease Information Swicki (http://addisons-disease-information-swicki.eurekster.com), který je zaměřen na problematiku Addisonovy choroby.

4. Posilování **komunikace**, **spolupráce** a **sdílení** informací v rámci **Community 2.0** vede ke vzniku řady nových, netradičních zdrojů informací. To otevírá nové možnosti vyhledávání odborných a vědeckých informací mimo tradiční profesionální informační zdroje, v prostředí online komunit, na webových stránkách jednotlivých odborníků (**blogy**) nebo profesních skupin (**social websites**, **social networks**, **user driven content websites, wikis**).
   Mezi systémy s uživateli řízeným obsahem (user driven content websites) patří například **BioWizard** (http://www.biowizard.com), jehož obsah vytváří komunita uživatelů výběrem článků z databáze PubMed. Systém umožňuje jejich prohlížení podle medicínských oborů a některých dalších biomedicínských specializací. Hlasováním je potom možné ovlivňovat pořadí v nabídce nejčtenějších článků. Uživatelé systému BioWizard tak mohou velmi rychle najít články nejvíce čtené a ceněné komunitou kolegů z celého světa.

Závěr

V oblasti informačních vyhledávacích systémů v posledních letech existuje velmi dynamický vývoj, který se projevuje širokou nabídkou obecně použitelných i pro medicínu a příbuzné obory specializovaných nástrojů. Aplikace uvedené v tomto sdělení jsou pouze ilustrací jednotlivých systémů a funkcí zaměřených na medicínu a související obory. Řada těchto systémů má více funkcí a mohla by být zahrnuta do více než jedné kategorie, od čehož však pro zachování přehlednosti bylo upuštěno. Provázanost a vzájemné souvislosti, stejně jako ilustrativní obrázky jednotlivých systémů však jsou obsahem posteru, který je grafickou alternativou tohoto článku.