Znalostní a referenční báze informačního analytika

PhDr. Jan ŽBIRKA
Tovek Partner, Praha
jan.zbirka@karneval.cz

INFORUM 2005: 11. konference o profesionálních informačních zdrojích
Praha, 24. - 26.5. 2005

Abstrakt. *Informační analytik se potřebuje v průběhu zadávání analýzy rychle a kvalifikovaně rozhodovat, zda bude schopen s disponibilními zdroji včas zpracovat analýzu požadované kvality. (Prakticky ve všech případech též řeší možnou míru kooperace s dalšími subjekty.) V rozhodnutí i zpracování mu může být nápomocna znalostní a referenční báze. Její znalostní struktury vznikají především zobecněním ontologií, taxonomií a dotazů. (Z nich pak vedou reference k značně heterogenním informačním zdrojům a jejich derivátům.) Existuje zde řada dílčích standardů, ale jejich integrace není triviální. Příspěvek na příkladech ukazuje typové problémy a jejich možná řešení.*

1 Úvod

V dnešním značně dynamickém světě potřebuje každý systém (nejen) na úrovni podniku, organizace či instituce pro řízení svého pohybu informace z vnitřního prostředí i relevantního okolí. Tyto informace pro rozhodování požaduje v přiměřené kvalitě, formě a čase, přičemž zejména kvalita a čas jsou hlavní kritéria působící proti sobě.

Tak jak se zvedá kvalita managementu od operativního k taktickému a strategickému řízení, tak se posouvá pozornost od základního zpracování dat k analýzám nejprve ve vnitřním a posléze i vnějším prostředí. Vnitřní data jsou obvykle omezenějšího rozsahu, lépe strukturovaná, s převládajícími číselnými údaji, které jsou snadněji analyzovatelné. Vnější data jsou naproti tomu velmi velkého rozsahu, většinou zdánlivě nestrukturovaná, v převládající textové formě. Vnitřní analýzy dávají odpovědi zejména na otázky jak efektivně či neefektivně systém v jaké své části pracuje, zatímco vnější analýzy odhalují hlavně aktuální či potenciální hrozby nebo příležitosti.

Tento příspěvek se zaměřuje na oblast analýzy vnějšího prostředí. V analytickém systému poukazuje na roztříštěnost dílčích technologií a na potřebu i možnosti jejich integrace, zejména pomocí znalostní báze a referenční báze, propojující znalosti s poznatky, dílčími informacemi či informačními zdroji.

2 Analytický systém

Analytický systém ve výše uvedeném významu můžeme chápat jednak v širším pojetí od vnějších informačních zdrojů, přes analytické pracoviště k uživatelům z vrcholového vedení. V užším pojetí můžeme chápat analytický systém jako klíčovou technologii analytického pracoviště.

2.1 Analytické pracoviště

Vstupní podsystém zajišťuje především vstupní zpracování informací z heterogenních zdrojů do homogenizované vnitřní podoby pro vyhledávání, analýzy a výstupní transformace. Bohužel často se zapomíná na funkci Akvizice. Měla by být alespoň na úrovni definování akvizičních pravidel.

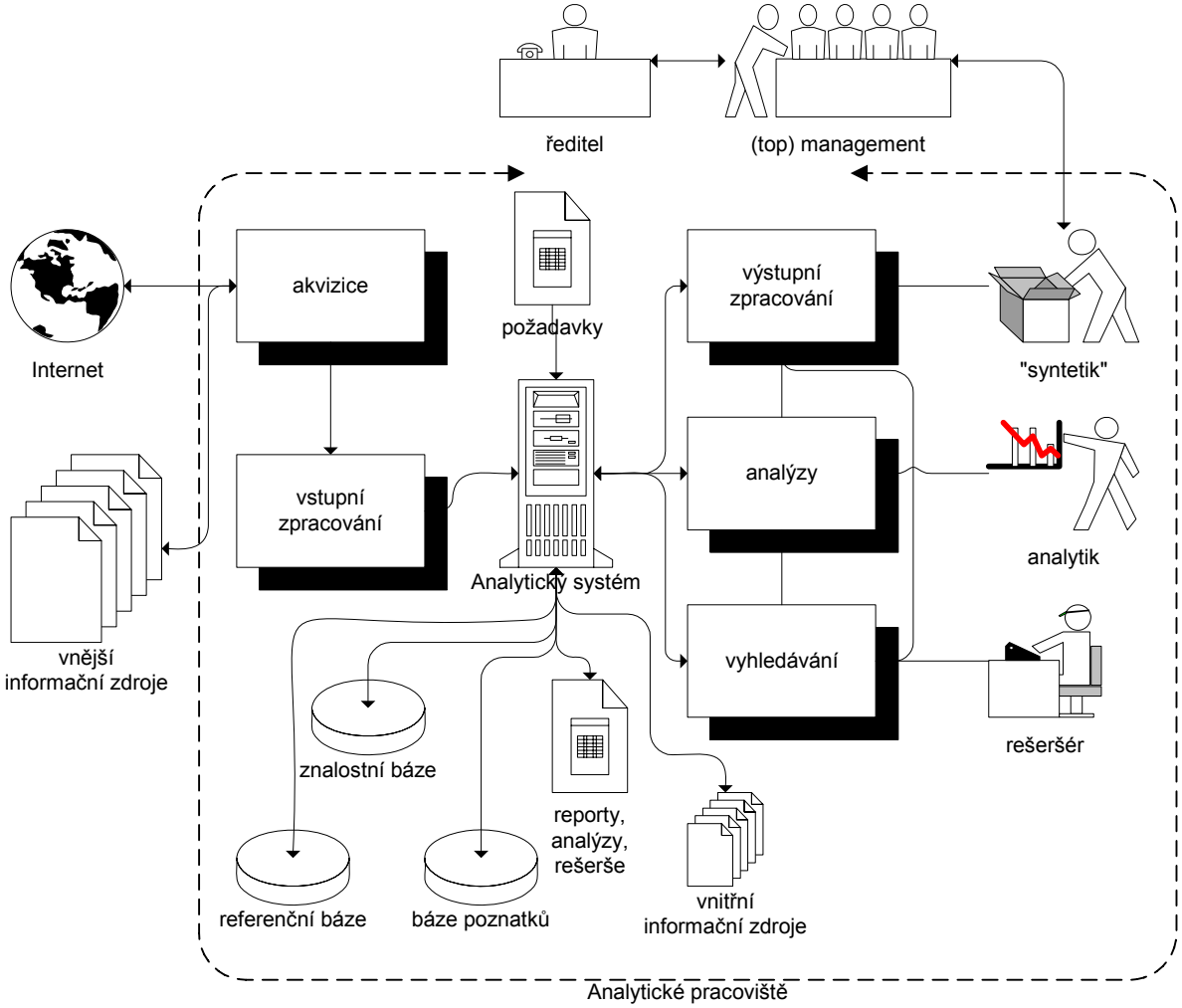
Jádrem analytického systému (v užším smyslu) je v současnosti fulltextový systém třetí generace (technologie Verity[1]) a systém vizuálních analýz (technologie I2[2]). Obě technologie jsou již delší dobu integrovány ve všech třech hlavních rovinách: dat, funkcí a rozhraní. Hlavním problémem je ale chybějící (a z hlediska výkonu velmi potřebná) integrace na úrovni znalostí do systémů již uložených (u fulltextu v podobě témat – Topiců, u vizuálních analýz v podobě vztahových diagramů).

Výstupní podsystém zahrnuje tři základní vrstvy: rešerše, analýzy a výstupní zpracování. Na zadání požadavku vrcholového vedení by měl reagovat reportem, jako podkladem pro kvalifikované rozhodnutí.

---

[1] http://www.verity.com/
[2] http://www.i2.co.uk/

Třem základním vrstvám výstupního podsystému odpovídají tři základní role: rešeršér, analytik, „syntetik“. Role rešeršéra a analytika jsou poměrně standardní s kompetencemi v pokročilých technologiích fulltextu a vizuálních analýz. Třetí role, pracovně nazvaná „syntetik“ potřebuje mít jak manažerské, tak informatické kompetence pro řízení zpracování analýz a závěrečné sestavení souhrnných výsledků do dílčích závěrů v reportu. Role rešeršéra a analytika lze snadno a s výhodou spojit v jedné fyzické osobě. Naopak role analytika a „syntetika“ je vhodnější pokrýt dvěma různými fyzickými osobami. Přechod mezi analytickým a syntetickým myšlením je u jedné osoby sice možný, ale vyžaduje přinejmenším určitou časovou prodlevu.



Obrázek 1– Analytický systém a analytické pracoviště

2.2 Znalostní báze

V analytickém systému vnějšího prostředí je v současnosti nejvíce znalostí uloženo ve fulltextových dotazech. Pro jednu firmu je typický hlavní dotaz (Topic) o objemu 200-500 uzlů a 50-100 doplňkových témat o objemu do 200 uzlů. Většina doplňkových témat jsou deriváty základního tématu s příslušným prohloubením. Tato témata a podtémata lze též využít pro vizuální analýzy, ale pouze jako uzly grafu, nikoli pro generování struktury požadovaného grafu. Témata určená pro vyhledávání jsou navíc informačně chudá, ovšem rozšiřování jejich struktury ve fulltextu by bylo spíše škodlivé. Rovněž údržba většího objemu témat uložených pouze v souborovém systému je obtížná.

Z výše uvedených důvodů jsme dospěli k potřebě vytvoření relativně nezávislé znalostní báze, použitelné nejen pro textové vyhledávání a vizuální analýzy, ale též pro komplexní práci analytického pracoviště od myšlenkových map, přes ontologie k jednotlivým i strukturovaným tematickým heslům, až po generování struktur požadovaných výstupních reportů.

2.3 Referenční báze

Referenční báze je oproti znalostní relativně jednoduchá. Jádro lze popsat řídkými maticemi čtyřrozměrného stavového prostoru, jehož dimenze jsou: uživatel, požadavek, téma (podtéma), dokument. Dokumenty nemusí být jen z vnějších či vnitřních informačních zdrojů, ale měly by to být i popisy požadavků, popisy témat, popisy reportů.

3 Datové struktury a jejich standardizace

Znalostní báze vzniká integrací proprietárních datových struktur (fulltext – Verity Query Language, vizuální analýzy I2) se standardizovanými (Topic Maps, DITA).

Rozsah a postup standardizace je vhodné sledovat na serveru OASIS[3] (Organization for the Advancement of Structured Information Standards)

Klíčovým problém je zde fakt, že pojem Téma-Topic používají minimálně 3 hlavní systémy v odlišném pojetí. To pramení z různých účelů těchto systémů.

3.1 Mapy námětů - Topic Maps

Mapy námětů (Topic Maps) jsou zde standardem nejsilnějším, neboť jsou již normou ISO. Umožňují znalosti reprezentovat sítí námětů – Topiců a vazbami na odpovídající informační zdroje. Pohyb po této síti je dynamický, vazby na zdroje statické.

Po své pražské přednášce The TAO of Topic Maps[4] mi Steve Pepper potvrdil moji hypotézu, že Topic Maps jsou silné v komplexním popisu tematické oblasti a pohybu do šířky, zatímco při pohybu do hloubky jsou velmi slabé a k přímému vyhledávání nepoužitelné. Shodli jsme se též na tom, že u Topiců ve Verity Query Language je tomu právě naopak a vzájemné propojení je velmi žádoucí.

Topic Maps jsou definovány na úrovni XML, resp. DTD, XSD.

3.2 Slovníky, tematické dokumenty – DITA

Architektura DITA[5] (Darwin Information Typing Architecture) je určena k vytváření tematicky orientovaných, vícenásobně použitelných dokumentů. Podporuje též proces zvaný specializace, určený ke kombinování a rozšiřování typů dokumentů.

Hlavním strukturním prvkem je zde opět téma – Topic. Kromě faktů může obsahovat další podtémata též typu Topic.

DITA je definována na úrovni XML, resp. DTD, XSD.

3.3 Dotazy- Verity Query Language

Dotazy (Topic) v syntaxi Verity Query Language (VQL) jsou určeny pro fulltextové vyhledávání a označování relevantních výrazů ve vyhledaných textech. Dobře konstruovaný (víceúrovňová struktura, dobré vyvážení) dotaz VQL má rovněž analytickou funkci. Lze jím tedy měřit texty a přímo převádět výsledky do grafů.

Jak bylo uvedeno výše u Topic Maps je VQL dobrý při postupu do hloubky a k analýzám, ovšem není vhodný pro komplexní popis tematické oblasti. I když umožňuje k tématům připojovat jednoduché poznámky, není použitelný ke strukturovanému záznamu faktů, tak jako DITA.

Dotaz VQL je definován v proprietárním formátu OTL, nikoli na úrovni XML, resp. DTD, XSD.

3.4 Vizuální analýzy - Analyst' s Notebook

Pro vizuální analýzy je podstatná možnost importovat části znalostních struktur (jednotlivá témata příp. jim odpovídající fulltextové dotazy, vzájemné vazby, fakta) jako podklady před zahájením analýz.

Analyst' s Notebook umožňuje import dat ve formátu XML.

---

[3] http://www.oasis-open.org/
[4] http://www.ontopia.net/topicmaps/materials/tao.html
[5] http://xml.coverpages.org/dita.html

4 Závěr – možnosti integrace systémů

K efektivní práci analytického pracoviště je potřebné, ale i možné oddělit znalosti uložené v jednotlivých systémech od jejich proprietárních formátů a uložit je ve společné znalostní bázi.

Návrh struktury této znalostní báze a výměna dat mezi systémy je vhodné založit na formátech XSD, resp. XML. Většina klíčových technologií toto přímo podporuje.

Pro Verity Query Language je potřeba schéma XSD připravit. Přitom jsou možné dva různé přístupy, jeden více přímo podporující pořizování a údržbu dotazů, druhý vhodnější pro integraci a výměnu dat. Lze uvažovat i o použití obou schémat v různých částech systému a jejich konverzi.

K analýzám a návrhu schémat XSD byl použit XML editor XML Spy[6], k definicím vzájemných transformací potom vizuální integrační editor Map Force[7].

Formát XML je vhodný pro výměnu dat, nikoli pro uložení znalostí a jejich správu. Proto je nyní testována postrelační databáze Caché[8]. Její objektové uložení dat odpovídá struktuře analytických znalostí. Nativní podpora vícerozměrných řídkých matic zase přímo odpovídá požadavkům na referenční bázi.

K přípravě schématu XSD pro výstupní reporty a pro definici výstupních transformací XSL byl opět použit editor XML Spy.

[6] http://www.altova.com/products_ide.html
[7] http://www.altova.com/products_mapforce.html
[8] http://www.intersystems.cz/cache/index.html