Google Knowledge Graph: things, not strings

ZUZANA PELIKÁNOVÁ

prosinec 2014

OBSAH

1 ÚVOD

Rozšíření chytrých telefonů, tabletů a dalších způsobilo, že lidé chtějí při vyhledávání rychle získat přesné a stručné informace, a to bez zdlouhavého hledání – to znamená, že vyhledávač má ideálně udělat všechnu práci za ně. Google má jako čelný představitel v oblasti vyhledávačů na rozšířeném, atraktivním a nově pojatém vyhledávání silnou motivaci; nejvýznamnějším nástrojem, který pro tyto účely používá, se nazývá Google Knowledge Graph.

2 OBECNĚ O KNOWLEDGE GRAPHU

Knowledge Graph je báze znalostí, kterou společnost Google poprvé představila v květnu 2012. Ze začátku si změněného vzhledu vyhledávače mohli všimnout pouze anglicky mluvící uživatelé, po několika měsících byli zapojeni i ostatní. Cílem Knowledge Graph je výrazně zlepšit vyhledávání, odstranit dvojznačné výrazy, umožnit uživatelům rozšířit si znalosti vyhledáním podobných či souvisejících témat a další. Způsob, jakým chce Google tohoto zlepšení dosáhnout, je patrný již z názvu - narozdíl od klasického pojetí vyhledávání, kdy jsou klíčová slova chápána pouze jako řetězce znaků, je nyní vytvořen rozsáhlý graf. Ten je tvořen uzly – jednotlivými entitami, například *Chuck Norris* – a hranami, které reprezentují vztahy mezi entitami (například *hrál v*, *narodil se v*). Entitě Chuck Norris odpovídají i entity muž, herec, narozený v roce 1940 a další.

3 MIKRODATA A SCHEMA.ORG

Rozšířené, strukturované vyhledávání je zajištěno různými způsoby. V této práci jsem se zaměřila na značkování webových stránek podle formátu mikrodat, které mohou využít i osoby, které chtějí, aby se obsah jejich webových stránek dostal do strukturovaných úryvků Vyhledávače Google (přestože to, v jakém pořadí se umístí mezi ostatními vyhledanými stránkami, samotné strukturování obsahu stránek neovlivní).

Mikrodata. Zápis informací pomocí mikrodat (rozšířeného zápisu HTML5, přesněji jeho specifikace) je jednoduchý způsob, jak je převést z webových stránek do strukturovanější a strojově čitelné podoby, například do znalostního grafu, a také je rozšířit o metadata. Extrakci takto strukturovaných dat poté provádí například crawlery.

Příklad několika základních atributů používaných při tvorbě mikrodat[1]:

itemscope	vytvoří objekt
itemprop	popisuje vlastnosti objektu
itemtype	URL slovníku, schéma (např. ze schema.org)

A velice jednoduchá ukázka kódu:

```
<div itemscope="http://schema.org/Person">
 <p>Jméno: <span itemprop="name">Zuzana Pelikánová</span>.</p>
 <p>Univerzita: <span itemprop="affiliation">Masarykova univerzita</span>.</p>
 <p>Věk: <span itemprop="age">21</span>.</p>
</div>
```

1 html.spec.whatwg.org/multipage/microdata.html

Schema.org. V roce 2011 vznikla iniciativa schema.org, na které spolupracoval Google, Microsoft Bing, Yahoo a další známé společnosti. Obsahuje slovník, jednotné schéma, jak mají webmasteři a další osoby postupovat při tvorbě strukturovaných dat. Obsahuje seznam povolených značek (například pro atributy itemtype, itemprop) při vytváření strukturovaného obsahu webových stránek. Přes existenci dalších standardů (například RDF) je toto schéma určené na práci s mikrodaty. Autoři se takto rozhodli zejména proto, že přestože je formát RDF poměrně rozšířený, není pro začátečníky zcela triviální na implementaci. Dalším rozšířeným standardem, který se zároveň je schopen naučit každý, jsou právě HTML5 mikrodata.

4 ZDROJE INFORMACÍ

Knowledge Graph čerpá informace především z Wikipedie, CIA World Factbook[2] a z dalších volně dostupných i licensovaných zdrojů na webu.

Freebase. Freebase je databáze, která tvoří významný zdroj vstupních dat pro Knowledge Graph. V současné době obsahuje přes 46 miliónů záznamů a asi 2,7 miliardy faktů[3], všechna data jsou strukturovaná a strojově čitelná (formát RDF). Databáze je automaticky extrahována z Wikipedie i dalších volně dostupných zdrojů, podstatnou část ale také ručně vytváří komunita přispěvatelů. Freebase spadá pod licensi Creative Commons (CC), takže její obsah je volně ke stažení[4] a je možné ho upravovat podle svého, zveřejňovat ve svých pracích (samozřejmě ocitovaný) apod.

Tato databáze je tvořena jako graf, kde uzly jsou označeny jako */type/object* a hrany definovány pomocí */type/link*. Základní prvek grafu Freebase je **záznam** (ve freebase označován jako topic). Záznam je totéž, co entita v Knowledge Graphu, tedy například *polární liška*, *hřeben*, *Leonardo da Vinci*. Záznamy obsahují **vlastnosti** (properties), které odpovídají vztahu HAS A - například *Praha* [=záznam] *má* [=vlastnost] *1,26 milionů obyvatel* [=konkrétní hodnota] nebo *Ian McKellen hrál ve filmu Hobit*. Záznamy jsou dále rozdělovány do jednotlivých kategorií, kterým se zde říká **typy**, types (cokoli od *člověk*, *ovoce*... po *umělecké směry v malířství*, *osobnosti 20. století* apod.). Typ odpovídá vztahu IS A. Každá entita může mít samozřejmě libovolný počet typů (Leonardo da Vinci byl člověk, muž, malíř, sochař, architekt...), přičemž každý typ je definován a nastaven odlišně. Typ herec uvede seznam filmů, ve kterých se objevil, typ společnost uvede jméno zakladatele, datum založení, sídlo, jméno současného ředitele společnosti atd. Další, vyšší patro tvoří tzv. **domény**, což jsou obecnější kategorie typu *film*, *hudba*, *ekonomika*, *sport* atd. Typům je přiděleno jednoznačné ID podle toho, do které domény patří. Typ herec například patří do domény film, jeho ID bude vypadat podobně jako například cesta k souboru: */film/actor*. Co se týče záznamů, ty jsou jednoznačně identifikovány pomocí *mid* - machine identifier, což je unikátní, strojově přidělovaný kód, identifikátor, který zajišťuje tzv. desambiguaci – zjednoznačněním významů slov. Ten je tvořen prefixem /m a řetězcem až sedmi znaků, tvořených písmeny i čísly; například mid hudební skupiny Daft Punk je /m/016j7m.

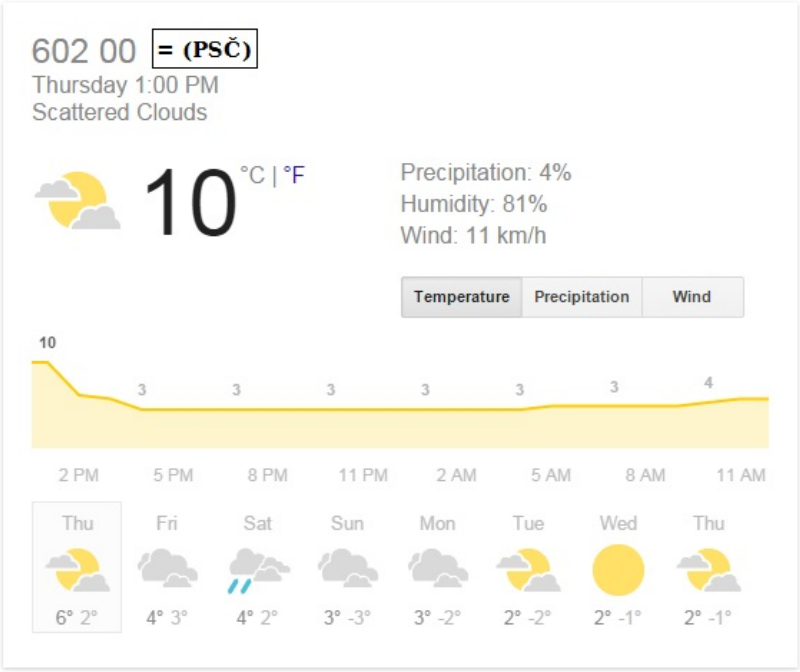
Hesla jsou velmi často dostupná v několika jazycích, stále jsou ale označena stejným identifikátorem. V nástrojové liště stránek freebase.com pod odkazem

2 www.cia.gov/library/publications/the-world-factbook
3 www.freebase.com
4 developers.google.com/freebase/data#freebase-rdf-dumps

I18n lze zjistit, jaký kód je přidělen jakému jazyku a pro které jazyky dané heslo zatím není vytvořeno.

5 VYUŽITÍ A PŘÍKLADY

Jedním z prvních způsobů, jak byl Knowledge Graph implementován do Vyhledávání Google, jsou odpovídací boxy. Ty jsou již v současnosti zaměřené na poměrně široké spektrum dotazů[5]. Od těch nejjednoduších typu *what's the time?*, *when is the next Easter?*, až po přesné odpovědi vázané na lokaci uživatele - *what is my IP address?*. Protože je Knowledge Graph implementován i pro mobilní aplikace, napadá mě, že toto rychlé zodpovídání dotazů mohou využít například turisté a další osoby, které se chtějí rychle dozvědět např. kdy mají otevřené památky, časy a místa odletů a příletů letadel, důležitá telefonní čísla (ambasáda a jiné), popřípadě jaká měna se kde používá, jaké je hlavní město země, atd.



Obrázek 1: Dotaz: What's the temperature today?

Informace zobrazené v odpovídacím boxu dříve nebyly nic jiného než jednotlivá data vytáhnutá z bočního pravého panelu (viz dále). V současnosti ale tzv. *answerbox* využívá i konkrétní webové stránky, například dotaz *social security tax rate* zodpoví odpovídacím boxem s částí článku z webových stránek časopisu Forbes. Je tedy možné data extrahovat i odjinud než z Knowledge Graphu, a do budoucna patří k velkým úkolům Googlu, aby zůstala kvalita úryvků v odpovídacím boxu vysoká. Uvedu příklad: na dotaz *what is May birthstone?* Google zobrazí odpovídací okno, kde kromě zvýrazněné stručné odpovědi (v tomto případě je talisman května smaragd) vypíše i stručný úryvek z webu; úryvek a zdroj se ale budou podstatně lišit podle toho, jak dotaz formulujeme. Některé dotazy mohou vést například k tomu, že Google zob-

5 V následující části používám dotazy v angličtině; pro češtinu jsou možnosti rozšířeného vyhledávání zatím bohužel dost limitované...

razí v boxu část reklamy z Ebay o koupi smaragdů, což je poměrně irelevantní a nežádoucí výsledek.

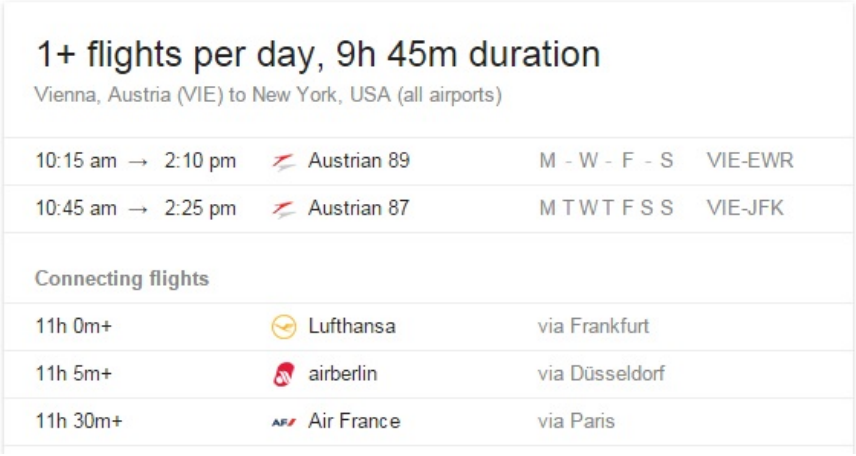
Obrázek 2: Panel, ze kterého jsou brána data pro odpovídací box – například jaká je adresa, kdy je otevřeno a další.

Nejviditelnější součást Knowledge Graphu ale představují boční panely, můžeme je také nazvat znalostní panely. Ty jsou umístěny na stránce výsledků vždy vpravo. Celkově dělají to, co by se dalo čekat – vypíší stručnou definici a pokud to jde, připojí k tomu obrázky a další informace, které se liší podle tématu. Panel hudebních interpretů například obsahuje členy kapely (i bývalé), významná alba, získaná ocenění, místo a datum vzniku a především nadcházející koncerty, jejich přesný čas a místo konání, to ale pouze tehdy, když se nachází v naší blízkosti (u některých interpretů to znamená Česká republika, u některých ale i celá Evropa). Zajímavý je znalostní panel pro ovoce, zeleninu a některé nápoje jako víno nebo pivo; ten uvádí kromě obvyklého také kompletní nutriční tabulku. Nevýhodou těchto panelů může být fakt, že většina informací pochází z Wikipedie či Freebase nehledě na to, zda jsou to opravdu ty nejvhodnější zdroje pro daná témata; pohled na ně může být tedy pouze jednostranný. Jedná se ovšem o nejjednodušší a nejlevnější řešení, a panely mají přeci jenom sloužit pouze ke stručné definici. Stejně jako u odpovídacích boxů Google začíná využívat i jiné webové zdroje (zatím ale pouze omezeně), které se ale nachází na konci panelu, kde nejsou dobře viditelné a mohou vypadat pouze jako reklama. Obrázky, které se nachází v panelu, jsou naopak většinou z jiných webových stránek, což je dobře.

Dalším rozšířením jsou takzvané *picture carousels*, karusely[6], které se zobrazují po zadání dotazů typu *well known actors* nebo *best Czech football teams*, nabízí tedy určitou skupinu entit (většinou kolem dvaceti) patřící do stejné

6 Doslovný český překlad na ”kolotoče” není úplně vhodný, proto v tomto textu zůstává výraz ”karusely”

skupiny. Google ale karusely nabízí pouze po zadání přesného řetězce slov nebo dalšími záhadnými způsoby, celkově by bylo lepší případy, pro které je tento způsob reprezentace vhodný, sjednotit.



Obrázek 3: Dotaz: flights to New York

6 SOUČASNÝ STAV A ZHODNOCENÍ

V průběhu roku 2014 byla data v Knowledge Graphu stále výrazně rozšiřována, (byla například přidána většina dat o počítačových hrách); v září oznámil Google ve své zprávě[7] rozšíření úryvků, které se zobrazují pod názvy vyhledaných stránek. Tato data získává Knowledge Graph nově i z tabulek. Jak uvádí zpráva, Google využívá k oddělení užitečných dat v tabulkách strojové učení, poté zvláštním algoritmem setřídí data podle relevance a kvality a rozhodne, které z nich budou zobrazeny. Důležitou (relativní) novinkou je znalostní báze s názvem Google Knowledge Vault ohlášená v srpnu 2014, která zahrnuje kolem 1,6 miliard faktů. Ta oproti Graphu shromažďuje data z celého webu, i z nejméně důvěryhodných zdrojů, a ta poté třídí opět pomocí strojového učení. V budoucnu se dá očekávat spíše rozšiřování Knowledge Vaultu, který už teď obsahuje i většinu dat Graphu.

Co se týká Knowledge Graphu obecně, hlavní nedostatek vidím už v základní myšlence, že znalosti = pasivní slovníkové informace z Wikipedie a jí podobných; znalosti jsou podle mého něco minimálně o stupeň víc. Nechci zde kritizovat úroveň informací na Wikipedii, pouze myslím, že moderní znalostní báze by měly být pojaty jinak, popřípadě vycházet z různorodějších zdrojů.

7 www.googleresearch.blogspot.cz/2014/09/introducing-structured-snippets-now.html

LITERATURA

[1] PUJARA, Jay, Hui MIAO, Lise GETOOR a William COHEN. Knowledge Graph Identification. In: *Lecture Notes in Computer Science*, s. 542. DOI: 10.1007/978-3-642-41335-3_34.

[2] SINGHAL, Amit: *Introducing the Knowledge Graph: Things, Not Strings.* Google Official Blog [online]. 2012 [cit. 2014-12-03]. Dostupné z: http://goo.gl/zivFV.

[3] TAYLOR, Jamie. *Introducing the Freebase RDF service* [online]. 2008 [cit. 2014-12-03]. Dostupné z: web.archive.org/web/20120516075431/http://blog.freebase.com/2008/10/30/introducing_the_rdf_service

[4] *Freebase* [online]. 2014 [cit. 2014-12-3]. Dostupné z: www.freebase.com

[5] *Schema.org* [online]. 2011 [cit. 2014-12-3]. Dostupné z: www.schema.org

[6] SCHWARZ, Ben. *HTML Living Standards* [online]. 2014 [cit. 2014-12-3]. Dostupné z: https://developers.whatwg.org

[7] CORTES, Corinna et al. *Introducing Structured Snippets, now a part of Google Web Search.* Google Research Blog [online]. 2014 [cit. 2014-12-03]. Dostupné z: www.googleresearch.blogspot.cz/2014/09/introducing-structured-snippets-now.html

METADATA V DC

dc:title: Google Knowledge Graph: things, not strings
dc:creator: Zuzana Pelikánová
dc:subject: sémantické vyhledávání, mikrodata, strukturované vyhledávání, báze znalostí
dc:date: 4. 12. 2014
dc:description: Tato esej se zabývá znalostní bází Google Knowledge Graph. Obsah tvoří popis znalostního grafu, způsob, jak se do něj extrahují data z běžné webové stránky, popis zdrojů pro znalostní graf, možnosti jeho využití běžným uživatelem a současný stav projektu.
dc:type: text
dc:language: cs