Nástroje Google. 5. Google Image Search

Miroslav Bartošek, ÚVT MU

Služba na vyhledávání obrázků na Internetu Google Image Search je po webovém vyhledávači druhou nejoblíbenější službou Google. A také jednou z nejstarších – uvedena byla již v roce 2001 (samotný webový vyhledávač začal do povědomí širší veřejnosti pronikat v letech 1999-2000).

1 Co to je

Služba Google Image Search (`http://images.google.com`, v české verzi `http://images.google.cz`) vyhledává obrázky na webu. Jde o obrázky jakéhokoliv typu – fotografie, kresby, schémata, grafy, klipart. V naprosté většině případů jde o obrázky, které jsou součástí textových webových stránek. Proto je služba těsně propojena s webovým vyhledávačem. Samotné obrázky jsou na webu identifikovány podle typu souboru (jpg, gif, png, bmp aj.). Vyhledávání v databázi obrázků je realizováno jako textové vyhledávání, kdy slova zadaná v dotazu jsou hledána v názvech souborů, textových popiscích hypertextových odkazů (atribut ALT), a hlavně v textu v okolí obrázku (je jedno, jestli jde o popisek obrázku nebo prostě jen text nacházející se na stránce, na níž je obrázek umístěn). Jako výsledek dotazu je zobrazena stránka vyhledaných obrázků v podobě *náhledů* (zmenšených obrázků – anglicky thumbnails); kliknutím na zvolený náhled se otevře stránka se dvěma rámy – v horním rámu je umístěn náhled obrázku s odkazem na obrázek v plné velikosti, ve spodním rámu je zobrazena celá webová stránka, na které je obrázek umístěn – uživatel má tak ihned k dispozici celý kontext, v němž se obrázek na webu nachází.

Formulář pro zadání vyhledávacího dotazu je zcela v googlovském duchu. Základní formulář využívaný většinou uživatelů je tvořen jediným políčkem pro zadání dotazu; formulář pro pokročilé vyhledávání umožňuje zpřesnit vyhledávací dotaz zadáním požadovaného typu obrázku (fotografie, klipart, kresba, tváře, obrázek ze zpravodajství), velikosti obrázku, typu souboru, barevnosti, webového zdroje a mírou filtrace pornografického obsahu (tzv. SafeSearch).

Protože veškeré zpracování – od vytváření databáze obrázků až po vyhledávání – je zcela automatizované, trpí služba obdobnými nešvary jako jiné vyhledávací služby Google: vyhledávání bývá někdy i značně nepřesné (o to více, že systém neumí hledat přímo v obrázcích, ale hledá v okolním textu, který nemusí se samotným obrázkem vůbec souviset); uživatel je obvykle zavalen velkým množstvím výsledků, z nichž některé nejsou vůbec relevantní; nejsou k dispozici žádné údaje o tom, co všechno a z jakých zdrojů vlastně databáze obrázků obsahuje. Na druhou stranu je však třeba říci, že ve většině případů tyto nedostatky uživatelům zase až tak nevadí. Uživatelé typicky nehledají konkrétní obrázek, ale „nějaký vhodný“ obrázek na dané téma. V takových případech není rozptyl výsledků na škodu, někdy může být dokonce i užitečný. Díky tomu, že vizuální vnímání je u lidí velmi efektivní, dokáže člověk vyhodnocovat obrázky (oddělovat zrno od plev) velmi rychle, a to i ve velmi rozsáhlé množině výsledků; na rozdíl od textových dokumentů to prostě „vidí hned“. Tradičním bonusem, obdobně jako u ostatních vyhledávacích služeb Google, je obrovský rozsah databáze (odhady hovoří o několika miliardách obrázků) pokrývající prakticky jakékoliv téma, velmi rychlé vyhledávání, snadnost použití a bezplatné využívání kýmkoliv, kdykoliv a odkudkoliv.

2 Různá vylepšení

Služba Google Image Search je od svého uvedení v roce 2001 průběžně zdokonalována. Vylepšení se týkají vyhledávání, řazení výsledků i obsahu vlastní databáze obrázků. Mnohá z těchto vylepšení zaznamenali uživatelé teprve v nedávné době. Některá vylepšení se opírají o lepší zpracování doprovodných textů a dostupných metadat, nověji se čím dál častěji objevují (zatím sice drobná leč přesto užitečná) vylepšení vycházející z automatizované analýzy samotného obrazu – rozpoznávání vzorů a podobnostních charakteristik či detekce objektů. Uveďme si aspoň ta nejzajímavější vylepšení (některé z nich nemusí být

implementována v české mutaci vyhledávače, doporučujeme přepnout na anglickou verzi).

2.1 Ruční klasifikace obrázků

K těm starším „technologiím“ pro zlepšování kvality vyhledávání patří tzv. *Google Image Laberer* `http://images.google.com/imagelabeler/`. Jedná se o zapojení uživatelů do klasifikace (popisování) obrázků formou zábavné hry. V každém kole, které trvá dvě minuty, je uživatel náhodně spárován s jiným hráčem někde v světě (bez možnosti vzájemné domluvy), a oběma jsou ukazovány postupně stejné obrázky; ke každému obrázku se snaží přiřadit co nejvíce hesel popisujících daný obrázek. V případě, že se hráči trefí u obrázku do stejného hesla, přičte se jim určitý počet bodů (za obecnější heslo je méně bodů než za konkrétnější). Cílem je získat co nejvíce bodů – ať již v jednorázové hře nebo v dlouhodobé soutěži mezi uživateli. Google tímto způsobem získává metadata sloužící pro přesnější vyhledávání obrázků.

2.2 Filtrování podle velikosti

Pokud uživatel potřebuje obrázek konkrétní velikosti nebo prostě jen obrázky co nejkvalitnější (největší), může využít některý z filtrů na velikost obrázků. Nejjednodušší variantou je zvolit na stránce vyhledaných obrázků některou z předdefinovaných velikostí (extra-velké, velké, střední, malé obrázky) a systém automaticky zúží množinu vyhledaných obrázků jen na obrázky požadované velikosti. Další možností je zadat přesnou velikost obrázku v pixelech do formuláře pokročilého (rozšířeného) vyhledávání nebo ji zapsat přímo do políčka jednoduchého vyhledávání; příklad: „imagesize:640x480 pes“.

2.3 Filtrování na základě barev

Tento filtr umožní vyhledat obrázky podle převládající barvy. Opět ho lze použít několika způsoby: na stránce s vyhledanými výsledky lze rozbalit barevnou paletu a z ní vybrat požadovaný barevný tón; druhou možností je přidat do URL výsledků vyhledávání frázi imgcolor=barva, například `http://images.google.cz/images?q=pes&imgcolor=yellow`.

2.4 Filtrování podle formátu

Pouze ve formuláři pro pokročilé vyhledávání lze zadat vyhledávání obrázků pouze v zadaném grafickém formátu. Na výběr jsou možnosti jpg, gif, png a bmp.

2.5 Filtrování na základě obsahu

Vyhledávání obrázků podle jejich obsahu využívá filtry dvou rozdílných typů – jeden typ slouží k volbě různých googlovských databází (např. filtr *news content* zobrazuje pouze obrázky ze zpravodajství), druhým typem jsou filtry založené na analýze obrazu (např. filtr *Faces* nebo filtr *Photo content*). Aktuálně jsou ve formuláři pro pokročilé vyhledávání nebo dokonce i přímo v uživatelském rozhraní na stránce výsledků k dispozici následující filtry obsahu:

- News content – obrázky ze zpravodajství z posledního měsíce (30 dnů),
- Faces – obrázky obsahující obličeje,
- Photo content – fotografie,
- Clipart – klipart,
- Line Drawings – kresby.

Google pracuje intenzivně na dalším rozšíření a využití obsahových filtrů. Jednou z možností je rozšířit detekci obličejů na rozpoznávání konkrétních tváří, což by podstatně zkvalitnilo vyhledávání obsahující jména osob. Řadu možností nabízí automatizovaná detekce různých objektů v obraze. Z jiného soudku je zase například využití EXIF metadat u digitálních fotografií (datum pořízení fotografie, geografická lokace atd.)

2.6 Řazení obrázků ve výsledcích

Pro každý dotaz zobrazí Google Image Search nejvýše 1 000 výsledků. Na jednu stranu je to hodně, a je proto důležité seřadit výsledky tak, aby nejrelevantnější obrázky byly zařazeny v popředí a uživatel je nepřehlédl. Na druhou stranu může pevně nastavený limit obrázků znamenat, že při nedostatečně kvalitním řadicím algoritmu mohou být některé relevantní obrázky uživatelům fakticky nedostupné. V každém případě je pořadí obrázků ve výsledcích velmi důležité. U klasického webového vyhledávače Google jsou

výsledky řazeny podle PageRank algoritmu zohledňujícího mimo jiné množství odkazů vedoucích z webu na danou webovou stránku a jejich váhu. Na rozdíl od webových stránek neposkytuje však PageRank u obrázků vždy dostatečně kvalitní řazení. Google proto pracuje na vylepšení. Jednou z vyvíjených technologií je tzv. *VisualRank* [3]; ten analyzuje obrázky z hlediska jejich vizuální podobnosti a s vypočtenými podobnostmi následně zachází jako s pravděpodobnostními vizuálními hyperlinky, které lze dále zpracovávat technologií PageRank. Podle tvrzeni Google dosahuje řazení s využitím VisualRank výrazně lepších výsledků než současné algoritmy používané v Google Image Search.

2.7 Podobné obrázky

Nejžhavější novinkou ve vyhledávání obrázků je funkce *Google Similar Images* `http://similar-images.googlelabs.com/`, která byla představena koncem dubna tohoto roku jako experimentální aplikace pro vývojáře a fanoušky v rámci laboratoří Google (`http://www.googlelabs.com`). Tato aplikace umožňuje vyhledat podobné obrázky ke zvolenému obrázku na základě analýzy obrazu (zřejmě s využitím algoritmu VisualRank – viz výše), nikoliv tedy na základě textového vyhledávání. Funguje následujícím způsobem: do vyhledávacího políčka zadáme dotaz pro vyhledání sady obrázků klasickým způsobem; pod náhledy (všech nebo jen některých) vyhledaných obrázků se objeví odkaz *Similar images*. Kliknutím na tento odkaz se zobrazí obrázky, které jsou vizuálně podobné zvolenému obrázku. Aplikace neumožňuje provádět analýzu obrazů v reálném čase (tj. neumožňuje uživateli nahrát jeho vlastní obrázek a k němu hledal podobné), navíc zdaleka ne pro všechny obrázky v databázi Google je funkce podobných obrázků nabízena (srovnejme například výsledky dotazu „ant“ s výsledky dotazu „mravenec“ – zatímco v prvním případě nabízí odkaz „Similar images“ většina vyhledaných obrázků, ve druhém případě naopak téměř žádný). I když má tato aplikace k dokonalosti ještě daleko, i tak je docela zajímavá a stojí za vyzkoušení.

2.8 Fotografie magazínu Life

V oblasti rozšiřování databáze obrázků Google je pozoruhodným počinem zařazení profesionálních reportážních fotografií z archívu časopisu Life. Celý archív, obsahující přes 10 miliónů vysoce kvalitních fotografií a rytin od více než stovky autorů, jdoucí až do roku 1750, Google skenuje ve vysokém rozlišení a digitalizované fotografie spolu s jejich popisy postupně včleňuje do své databáze obrázků (projekt byl zveřejněn v listopadu 2008 [4], a v té době bylo již zařazeno do databáze Google okolo dvou miliónů historických fotografií). Při běžném hledání v Google Image Search jsou tyto fotografie rozptýleny mezi ostatní výsledky (náhledy fotografií na stránce výsledků hledání jsou označeny tagem LIFE). Pokud chceme vyhledávat přímo jen fotografie z archivu Life, je možné buď přidat frázi „source:life“ do vyhledávacího dotazu (např. „vietnam war source:life“), nebo jít přímo na stránky archívu Life na adrese `http://images.google.com/hosted/life`.

3 Konkurenční systémy

Google Image Search není jediná služba pro vyhledávání obrázků na webu, i když je to zřejmě služba nejznámější a nejužívanější. Podle [5] patří do „velké trojky“ v této kategorii ještě vyhledávače Yahoo-images `http://images.search.yahoo.com/` a Microsoft Live-Search `http://www.live.com/?scope=images`. Kromě toho existuje řada „menších“ vyhledávačů, které nemají takové pokrytí, často ale nabízí zajímavé pokročilé funkce, které velké vyhledávače ještě neumí (jmenujme alespoň švédský Picsearch `http://www.picsearch.com/` a francouzský Exalead `http://www.exalead.fr/image/`).

Pro čtenáře z MU je jistě zajímavá informace, že na poli strojové analýzy obrazu a podobnostního hledání obrázků dosáhl pozoruhodných výsledků výzkumný tým z Fakulty informatiky MU vedený prof. Zezulou, viz článek [6]. Jejich experimentální systém MUFIN (Multi-feature Indexing Network), pracující aktuálně nad databází 100 miliónů fotografií ze služby Flickr, si ostatně čtenáři mohou sami vyzkoušet na `http://mufin.`

fi.muni.cz/imgsearch/ – a porovnat ho například s výše uvedeným systémem Google Similar Images.

3.1 Závěr

Služba Google Image Search pro vyhledávání obrázků na Internetu je velmi populární a hojně využívaný nástroj, který neslouží jen k vyhledávání fotografií celebrit, ale i k seriózní práci a vzdělávání, zejména při vizualizaci pro nejrůznější účely. Obrovská databáze obrázků, snadné používání a velmi rychlá odezva jsou přednosti, které dokáží překrýt i občasné problémy s přesností výsledků vycházející z nedostatků textového vyhledávání. Pokroky v oblasti automatizované analýzy obrazu a na ně navazující pokročilé vyhledávací funkce naznačují, že v dohledné době můžeme očekávat v oblasti vyhledávání obrázků mnohá zajímavá překvapení.

Literatura

[1] Google Help. Google Image Search: Image Search Help. http://www.google.com/support/websearch/bin/answer.py?hl=en&answer=138524

[2] *Image Search*. Google Operating System. Unofficial news and tips about Google. http://googlesystem.blogspot.com/search/label/ImageSearch

[3] Y. Jing, S. Baluja. *PageRank for Product Image Search*. WWW 2008. Beijing, China, April 21-25, 2008. http://www.www2008.org/papers/pdf/p307-jingA.pdf

[4] *LIFE Photo Archive available on Google Image Search*. The Official Google Blog. http://googleblog.blogspot.com/2008/11/life-photo-archive-available-on-google.html

[5] J. Hargreaves. *Review of Image Search Engines*. JISC Digital Media Review, 21.2.2008 http://www.jiscdigitalmedia.ac.uk/stillimages/advice/review-of-image-search-engines/

[6] P. Kohoutková. *MUFIN: buďte v obraze!* Ikaros [online]. 2009, roč. 13, č. 3. ISSN 1212-5075. http://www.ikaros.cz/node/5301 □