Extrakce informací z emailů typu Call for papers

Martin Dostal, Lubomír Krčmář, Karel Ježek

Katedra informatiky a výpočetní techniky, Fakulta aplikovaných věd
Západočeská univerzita v Plzni
Univerzitní 22, 306 14 Plzeň, Česká republika
{madostal,lkrcmar,jezek_ka}@kiv.zcu.cz

Abstrakt *Každý den dostává většina akademických pracovníků množství emailů typu Call for papers (CFP), neboli oznámení o konferencích. Organizace těchto emailů zabírá stále více času a aktualizace údajů v kalendáři je často více než náročná. V rámci tohoto článku typu Work in progress bychom chtěli představit systém pro extrakci informací z těchto oznámení. Pro dolování informací využíváme sadu jednoduchých, ale efektivních technik v nevšedním pojetí. Jde např. o extrakci informací (EI) na základě n-gramů, nebo s využitím vlastní implementace webového rozhraní k populárnímu nástroji GATE [1].*

1 Úvod

Cílem práce je vytvoření nástroje, který umožní získat lepší přehled o budoucích konferencích. Jedná se hlavně o důležité datumy, témata, místo a čas konání konference. Nástroj bude nejdříve využit pro potřeby konkrétních výzkumných skupin, později je předpokládáno jeho veřejné zpřístupnění.

Hlavní motivací práce je fakt, že v dnešní době je většina konferencí oznamována pouze emailem a není možné získat celkový přehled o jednotlivých událostech. Často se stává, že je některá zajímavá konference opomenuta např. z důvodu promeškání důležitého termínu, nebo přehlédnutí zajímavého tématu. Tento nástroj se snaží nejen získávat informace z vlastních oznámení, ale navíc dohledává další informace na webových stránkách oznámených konferencí, nebo přímo sbírá samostatná oznámení z prostředí Webu. Tato úloha kombinuje problematiku extrakce a klasifikace informací z prostého textu i Webu s technikami pro sdílení informací v rámci sociálních sítí s využitím principů Webu 2.0. Přístup k seznamu konferencí je pro uživatele umožněn přes webové rozhraní a uživatelé mohou využívat pokročilého filtrování a štítkování konferencí pro sebe i své sociální skupiny. Již v průběhu nastavení omezujících podmínek se uživateli zobrazuje počet vybraných konferencí. Této vlastnosti je dosaženo s využitím interaktivního webového rozhraní založeného na technologii AJAX. Tato webová aplikace využívá existující knihovnu jQuery [2] pro podporu technik Webu 2.0.

1.1 Obsah článku

V tomto článku budou vysvětleny použité existující i navržené metody pro extrakci a klasifikaci informací. Dále budou prezentovány předběžné výsledky týkající se přesnosti a úplnosti extrahovaných dat.

Metody pro extrakci a klasifikaci informací budou představeny zvlášť z důvodu přehlednosti a využití paralelních nástrojů řešících tyto úlohy odděleně. Při extrakci informací se budeme zabývat pouze výběrem zajímavých bloků textu a v rámci klasifikace budeme řešit jejich kvalitu, hodnotu a správnost vzhledem k celému informačnímu zdroji. Výběr zajímavých částí textu provádíme s využitím vlastní webové služby založené na nástroji GATE a paralelně i s využitím extrakce založené na n-gramech. Extrakce informací založená na n-gramech je velice rychlá a poměrně spolehlivá, neobsahuje však tolik možností jako nástroj Gate online. Gate online navíc umožňuje využití gazetteerů, neboli seznamů pojmenovaných entit jako např. anglické názvy měst a států. Oproti extrakci založené na n-gramech dosahuje vyšší úplnosti označených dat, avšak nižší přesnosti. Detailnější porovnání bude provedeno v rámci popisu metod pro extrakci informací.

2 Gate online

Komponenta Gate online, jak její název napovídá, využívá nástroje GATE [1] a funguje jako webová služba. GATE je open source software, který je použitelný k řešení téměř všech úloh týkajících se zpracování přirozeného jazyka. GATE je obecný a univerzální nástroj složený z mnoha pluginů a nabízející Java API pro vytváření aplikací. Nástroj GATE je používán početnou komunitou především pro extrakci pojmenovaných entit. Konkrétními příklady jsou i systém MACE [3] nebo metoda BEIRA [4].

Gate online je dalším nástrojem využívajícím GATE Java API. Na rozdíl od jiných aplikací, které se používají k řešení specifických úloh, je Gate online nástrojem více obecným. K obecnosti přispívá využití základních dvou pluginů, kterými jsou ANNIE a Montreal Transducer. Gate online vzniká jako znovupoužitelný nástroj postavený na Java GATE API pro extrakci informací z textů.

2.1 Způsob využití Gate online

Komponenta Gate online je nyní v aplikaci pro analýzu emailů typu CFP využívána přes HTTP Post. Vstupem komponenty je analyzovaný text a parametry, mezi které patří

```
<Sentence gate:gateId="75">
  <Date gate:gateId="85" rule1="DateName" kind="date"
    rule2="DateOnlyFinal">14 July 2010</Date>
  <Location gate:gateId="86" rule1="Location1" locType="city"
    rule2="LocFinal">London</Location>
  ,
  <Location gate:gateId="79" rule1="Location1" locType="country"
    rule2="LocFinal">United Kingdom</Location>
</Sentence>
```

Obrázek 1. Část XML výstupu získaného pomocí Gate online.

požadované či nežádoucí anotace. Jedním z možných vstupů je také množina skriptů napsaných v jazyce JAPE. Výstupem je XML dokument s tagy a jejich parametry, které odpovídají žádaným anotacím. Část ukázkového výstupu znázorňuje Obr. 1.

2.2 Gate online ve vývoji

Nástroj Gate online se postupně vyvíjí. Od verze GATE 5.1 není plugin Montreal Transducer dále podporován a je využívána jeho nová verze JCompiler. Gate online pracuje z technických důvodů se starší verzí pluginu. Způsob využití pluginu Montreal Transducer je nastíněn dále v textu.

V současné době je Gate online používán k extrakci anglicky psaných lokací a jmen osob z textů emailů. Extrakci umožňuje plugin ANNIE a seznamy klíčových slov (gazetteers). ANNIE dále umožňuje extrakci datumů zapsaných v anglickém formátu, český formát není podporován. Anotace českých datumů je realizována s využitím skriptů napsaných v jazyce JAPE.

K rozšíření možností webové služby byl proto integrován plugin Montreal Transducer. Tento plugin umožňuje překládat skripty napsané v jazyce JAPE. Gramatika jazyka JAPE není složitá a umožňuje pracovat s anotacemi jako s regulárními výrazy. Navíc lze v JAPE skriptech pro složitější operace s anotacemi používat i jazyk Java.

Plugin Montreal Transducer a v něm napsaný JAPE skript nyní umožňuje i extrakci českých typů datumů (např.: 22.10. 2009) nebo anotaci důležitých řetězců představujících typy událostí (např.: Paper submission).

Pomocí JAPE skriptů lze jednoduché anotace spojovat do složitějších a sémanticky významnějších. Pro případ anotace emailů typu CFP by bylo vhodné k datumům přiřazovat události, se kterými jsou spojené. Tato možnost není však ještě realizována, protože se v komponentě Gate online zatím z neznámých důvodů nedaří využít všechny možnosti jazyka JAPE. Konkrétně nefungují jazykové konstrukce „contains“ a „within“.

3 Extrakce informací z prostého textu

Jedním ze základních požadavků na vyvíjený nástroj je zvýšení komfortu při zpracování a plánování konferencí. Aby bylo možné tohoto cíle dosáhnout, je třeba provádět automatické zpracování získaných informací. Oznámení o konferenci můžeme získat z emailů, nebo z Webu. Ve většině případů se bude jednat o analýzu nestrukturovaných textových dat. Připravujeme i webového robota, který bude získávat oznámení z vybraných stránek na základě statistických metod i znalosti struktury stránky. Tyto metody budou však víceméně ojedinělé a nelze na ně spoléhat.

Přestože extrakce informací zahrnuje zároveň i klasifikaci, v našem případě bylo vhodné tyto dvě úlohy oddělit. Pod pojmem extrakce informací bude označováno nalezení zajímavých informací odpovídající šabloně, nebo vzoru. Problematika klasifikace textů bude v rámci této úlohy zaměřena na analýzu potenciálně zajímavých textů získaných v rámci extrakce informací. Jedná se o klasifikaci datumů, témat a dalších informací.

Toto rozdělení problému přináší možnost paralelizace výpočtu a rychlejší odezvu celého systému. Navíc se každá z těchto částí může v čase měnit aniž by omezila, nebo dokonce ohrozila provoz ostatních komponent.

3.1 Vliv předzpracování

Cílem předzpracování je usnadnění EI z prostého textu. Z tohoto důvodu je vhodné ze vstupního textu odstranit všechny html značky, abychom mohli používat naprosto stejných metod pro EI z emailů i webových stránek. Než tak můžeme učinit, je třeba získat z html značek použitelné informace. Jedná se např. o odkazy, nebo informace týkající se optické struktury textu. Jestliže nalezneme např. značku označující několik odřádkování nebo odstavec, je zřejmé, že se autor dokumentu snažil provést optické oddělení bloků textu. Tato informace pro nás může být důležitá. Tuto pozici v textu si označíme vlastní pomocnou značkou označující optické oddělení textových bloků. Nyní již lze bez problémů provést odstranění všech html značek, neboť podstatá informace z nich již byla získána.

Při analýze testovacího korpusu s emaily typu CFP bylo zjištěno, že většina textů je rozdělena do opticky oddělených textových bloků. Může to být dělení s využitím html značek, nebo optické oddělení s využitím např. znaků hvězdička, pomlčka apod. Každý optický blok obsahuje ve většině případů pouze jeden typ informace. Např. úvodní informace, důležité datumy, témata konference, kontaktní údaje a upřesňující informace o místě konání konference. Každý blok můžeme předběžně analyzovat a rozhodnout, zda obsahuje nějakou užitečnou informaci a odhadnout kterou. Pokud se nám podaří úspěšně označit blok textu obsahující důležité datumy, není třeba další datumy vyhledávat v ostatních blocích. Vlastní extrakci událostí s pomocí n-gramů tak můžeme provést pouze nad tímto jedním blokem, což nám velmi výrazně zvýší efektivitu celého systému. Výsledkem je analýza průměrně jen 1/5 původního textu. Blok obsahující témata konference je možné dále vy-

užít pro automatické štítkování a usnadnit tak třídění konferencí.

3.2 Využití n-gramů pro extrakci událostí

Za událost je v rámci této práce považována krátká textová informace doplněná o související datum. Extrakce důležitých událostí je prováděna nad textovým blokem zvoleným v rámci předzpracování. Tento blok obsahuje množství datumů a krátkých popisných textů. Nejvhodnější je zpracovávat textový blok v několika krocích. V rámci prvního kroku se provede hrubá klasifikace slov, která by mohla označovat den, měsíc, rok nebo jejich kombinaci. Ve druhém kroku se v rámci trigramů provede spojení těchto slov do jednoho celku a zároveň se provede převod na jednotný formát ve tvaru: YYYY-MM-DD. Tento tvar nám usnadní další zpracování a uložení do databáze. Nyní již zbývá krátké texty pouze spárovat s datumy. S využitím této techniky lze velice úspěšně vyextrahovat většinu datumů zapsaných v jednom z mnoha používaných zápisů včetně automatické opravy nejčastějších chyb.

3.3 Porovnání EI dle n-gramů a GATE

V případě extrakce událostí s využitím n-gramů je dosaženo velmi vysoké přesnosti i úplnosti získaných dat v rámci vybraného bloku. V případě webové služby Gate online je jako vstup využit celý dokument, nad kterým je prováděno automatické anotování. Tento přístup však neumožňuje označení vybraného bloku pro přesnější extrakci např. důležitých datumů, které se v tomto bloku pravděpodobně vyskytují. Webová služba Gate online anotuje všechny pravděpodobné výskyty datumů, již však neřeší jejich správnost ani převod na jednotný formát, což znesnadňuje další zpracování.

V rámci této úlohy je považována za důležitější přesnost získaných datumů, nikoliv úplnost. Dalším důvodem pro využití n-gramů pro extrakci datumů byla výrazně vyšší rychlost prováděné extrakce oproti službě Gate online. Služba Gate online se z tohoto důvodu využívá hlavně pro kontrolu a potvrzení datumů a specializuje se na složitější úlohy, které by nebylo možné jinak efektivně řešit. Jedná se např. o extrakci zeměpisných lokací, jmen osob apod.

4 Klasifikace údajů

V průběhu klasifikace dochází k výběru zajímavých údajů, které budou použity při vyhledávání, kategorizaci a štítkování konferencí. Získané údaje jsou zařazovány do vybraných tříd dle jejich ohodnocení určeného pro každou klasifikační třídu.

Nejdříve je třeba nadefinovat klasifikační třídy, vytvořit vzorec pro klasifikaci krátkých textů a zavést priority klasifikačních tříd.

4.1 Definice klasifikačních tříd

Volba klasifikačních tříd závisí na typech sledovaných informací. Může se jednat např. o informaci týkající se zaslání abstraktu, odeslání článku, přijetí příspěvku nebo konání konference. Údaj spadající do konkrétní klasifikační třídy může i nemusí být nalezen v rámci oznámení o konferenci. Nalezené údaje mohou často spadat do více klasifikačních tříd, což velmi znesnadňuje tuto úlohu. V průběhu klasifikační fáze je snaha o co nejpřesnější zařazení informace do vhodné klasifikační třídy. Pokud je textová informace rozpoznána a splňuje-li všechny omezující podmínky, je vždy zařazena právě do jedné klasifikační třídy. Možnost záměny je řešena až v rámci grafického uživatelského prostředí na základě priority a preferencí uživatele.

Každá klasifikační třída je definována množinou klíčových slov, která byla automaticky získána z bloku *důležitých datumů* v rámci bodu extrakce informací. Po odstranění stop slov byla klíčová slova seřazena dle jejich výskytu a nejfrekventovanější slova byla manuálně přiřazena klasifikačním třídám. Jedno klíčové slovo může být přiřazeno několika klasifikačním třídám s kladným i záporným ohodnocením. Kladné ohodnocení třídu potvrzuje, záporné vyvrací. Definice klasifikační třídy s využitím množiny klíčových slov se ukázala být mnohem vhodnější, než využití víceslovných frází. Díky tomuto přístupu dojde ke snížení časové i paměťové náročnosti klasifikačního algoritmu bez měřitelného vlivu na přesnost a úplnost extrahovaných dat. Využití frází dále nebylo možné realizovat z důvodu nedostupnosti dostatečně velkého korpusu, který by tyto fráze obsahoval. Náš současný korpus obsahuje cca 1000 CFP a způsob zápisu jednotlivých událostí se ukázal jako velmi různorodý.

4.2 Volba klasifikačního algoritmu

Pro potřeby této úlohy by mohlo být využito mnoho různých klasifikačních algoritmů jako např. rozhodovací stromy, kaskádové klasifikátory, nebo naivní Bayesovský klasifikátor.

Rozhodovací stromy se původně zdály jako nejlepší volba. Od tohoto řešení nás však odradila nutnost manuální, případně supervizované tvorby rozhodovacích stromů. V našem případě by se jednalo konkrétně o pravděpodobnostní rozhodovací strom. Při průchodu stromem se vybírají větve s maximálním pravděpodobnostním ohodnocením. Cílový stav je pak dán pouze cestou od kořene k listu. V případě využití této metody nedochází k úplnému průchodu rozhodovacího stromu a mohlo by dojít k opomenutí cesty s lepší pravděpodobností.

V rámci této úlohy bylo vhodnější využít funkci, která textu přiřadí ohodnocení pro všechny klasifikační třídy na základě všech nalezených údajů. Vybrána bude třída s nejlepším ohodnocením. Tomuto požadavku vyhovuje princip naivního Bayesovského klasifikátoru [5]. Pro tuto úlohu

je to však příliš mocný nástroj, který bude spotřebovávat zdroje bez odpovídajících výsledků. Tato úloha se zaměřuje pouze na klasifikaci událostí, tedy velmi krátkých textů. Z tohoto důvodu je zbytečné počítat podmíněnou pravděpodobnost slova na klasifikační třídě. Úplně si vystačíme s jeho existencí a celočíselným ohodnocením určujícím jeho význam. Pro tyto účely je však nutné klasifikační funkci upravit.

4.3 Volba klasifikační funkce

Tato úloha se zaměřuje na klasifikaci krátkých textů do předem stanovených tříd. Z tohoto důvodu lze stanovit následující kritéria pro klasifikační funkci:

1. Vytvoříme funkci vracející celočíselné ohodnocení textu pro každou klasifikační třídu. Celkové ohodnocení textu je dáno jako součet ohodnocení pro všechna vstupní slova. Každé slovo má určené ohodnocení dle jeho významu ke klasifikační třídě. Toto ohodnocení bylo definováno v rámci definice klasifikačních tříd. Pozitivní ohodnocení slova danou klasifikační třídu potvrzuje, negativní vyvrací.
2. Text zařadíme do třídy c s nejvyšším ohodnocením určeným dle vztahu (1), kde T je množina slov vstupního textu, s_i slovo z T a funkce v vrací ohodnocení slova dle jeho definice v rámci klasifikační třídy.

$$h(T|c) = \sum_{i=1}^{n} \nu(s_i|c). \quad (1)$$

Funkce ν z (1) je definována jako:

- k_1 pro $\forall s_i \in c$
- k_2 pro $\forall s_i \in E$, kde E je množina klíčových slov vylučujících třídu c
- k_3 pro $\forall s_i : \exists w_i \in c$, kde w_i je podřetězec s_i

Konstanty byly experimentálně nastaveny následovně: $k_1 = 1$, $k_1 = -10$, $k_3 = 0.5$. Konstanta k_3 charakterizuje pouze nalezení klíčového slova v podobě podřetězce analyzovaného slova. V rámci dalšího výzkumu lze předpokládat přesnější ohodnocení charakterizující význam slova pro klasifikační třídu.

Výsledný klasifikátor (2) splňuje všechna stanovená kritéria.

$$classify(T) = \arg\max_{c} \sum_{i=1}^{n} \nu(s_i|c). \quad (2)$$

4.4 Priority klasifikačních tříd

V průběhu klasifikace krátkých textů může poměrně často dojít ke stejnému ohodnocení dvou a více klasifikačních tříd. V tomto případě by klasifikátor zvolil první ze stejně ohodnocených tříd. Na základě počtu výskytů jednotlivých klasifikačních tříd však lze zvolit jejich prioritu, což tento problém odstraní. Výsledkem je volba priority tříd v tomto pořadí, počítáno od nejfrekventovanější třídy: událost odeslání článku, událost odeslání abstraktu, informace o konání workshopu atd.

Po zavedení priority tříd došlo ke zvýšení přesnosti až o 8%.

5 Dolování informací z Webu

Nedílnou součástí této aplikace bude i několik robotů, kteří budou chybějící informace dohledávat na Webu. Může se jednat o získávání informací přímo z oficiálních stránek konference, sběr oznámeních o konferenci např. z WikiCFP [6], nebo sběr dat z některého z populárních indexů – např. DBLP [7]. Robot pro extrakci informací z DBLP je již ve zkušebním provozu.

Pro sběr informací z indexu DBLP bude využito wrapperu, neboli znalosti struktury html stránky. Oznámení o konferenci bude z webu extrahováno s využitím jednoduché šablony založené na hledání např. nadpisu „Call for papers“ v kombinaci s některými klíčovými slovy jako např. abstrakt, článek, odeslání. Toto oznámení o konferenci bude uloženo do databáze MySql a dále zpracováváno stejnými technikami, jako by šlo o oznámení získané z textu emailu.

6 Průběžné výsledky

Uživatelům je umožněno provést potvrzení vyextrahovaných dat jedním kliknutím myši. V případě neshody mohou nesprávnou informaci opravit. Tyto údaje pak lze v budoucnu použít pro trénování systému. Systém není na těchto informacích závislý a funguje plně automaticky i bez manuální anotace. Vyhodnocení přesnosti a úplnosti extrakce se však bez manuální anotace neobejde, a proto byla tato funkce zpřístupněna běžným uživatelům. Při průběžném vyhodnocení přesnosti a úplnosti extrakce byla využita náhodná množina anotovaných konferencí obsahující přibližně 1/5 celého testovacího korpusu. Testovací korpus obsahuje přibližně 1200 emailů a z toho 1000 emailů je typu CFP. V této fázi nebylo možné z časových důvodů ručně anotovat celý korpus.

Úspěšnost extrakce informace o události je zanesena v tabulce č. 1 a vyjadřuje přesnost i úplnost extrahovaných dat. Tento údaj nevyjadřuje pouze správnost získaného datumu, ale zahrnuje i informaci zda je v daném emailu tento typ události obsažen či nikoliv. Úspěšnost vyjadřuje poměr správně vyextrahované informace (správný datum, nebo informace o nenalezení události) ku všem anotovaným emailům. V tabulce je uvedeno pouze několik ilustračních případů klasifikačních tříd.

Dále byla porovnávána úplnost nalezených událostí typu datum pouze v případě jejich existence v rámci CFP. Tato úplnost nalezení datumu je zanesena v tabulce č. 2.

Tabulka 1. Úspěšnost EI vybraných událostí

Událost	Úspěšnost
Odeslat abstrakt (datum)	92%
Odeslat článek (datum)	77%
Camera-ready verze (datum)	91%

Tabulka 2. Úplnost nalezených datumů

Událost	Úplnost
Odeslat abstrakt (datum)	93%
Odeslat článek (datum)	97%
Camera-ready verze (datum)	100%

V tabulce č. 3 byla zaznamenána přesnost zjištěných událostí typu datum pouze pro případy jejich zadání v rámci textu CFP. Přesnost extrakce třídy týkající se odeslání abstraktu je zkreslena relativně malým počtem anotovaných konferencí s touto událostí v testovacím korpusu. Lze očekávat, že přesnost extrakce třídy týkající se odeslání abstraktu bude na větším anotovaném korpusu o něco větší.

Tabulka 3. Přesnost nalezených datumů

Událost	Přesnost
Odeslat abstrakt (datum)	76%
Odeslat článek (datum)	77%
Camera-ready verze (datum)	93%

7 Další výzkum

Další výzkum se bude věnovat dvěma hlavním směrům: extrakci informací s využitím nástroje Gate online a pokročilým metodám štítkování v kombinaci s principy linked data.

7.1 Zdokonalování systému Gate online

Dalších plánů na rozšíření využití služby Gate online je několik. Jedním z nich je umožnění extrakce českých lokací a českých jmen z textů rozšířením Gate slovníků (gazetteers). Myšlenka, kterou se budeme také zabývat, je automatický převod datumů do jednotného formátu již v nástroji Gate online. V neposlední řadě zkoumáme další možnosti extrakce informací pomocí JAPE skriptů například ke spojování důležitých událostí s datumy jejich konání.

7.2 Napojení štítků na linked data

Budeme se zabývat dvěma metodami štítkování: automatické a manuální. Manuální štítkování by měli provádět sami uživatelé dle oblastí jejich zájmů. Automatické štítkování bude provádět robot na základě zaměření konference. Automatických štítků bude využito i pro označení geografických míst konání konferencí.

Štítky se pokusíme převádět na URI zdroje a napojovat na ostatní populární URI zdroje ze světa. Např. na entity z dbpedie [8]. Díky tomuto přístupu nám nic nebrání v dalším odvozování nad štítky, tvorbě synonym i hierarchické struktury štítků. Na základě tohoto přístupu bude vznikat množina RDF trojic, jejíž části mohou být modelovány s využitím ontologie, případně namapovány na některé existující ontologie např. z oblasti informatiky.

Poděkování

Tato práce byla částečně podporována z prostředků Národního Programu Výzkumu II, projekt 2C06009 (COT-SEWing).

Reference

1. General Architecture for Text Engineering“. URL: http://gate.ac.uk/
2. „jQuery: The Write Less, Do More, JavaScript Library”. URL: http://jquery.com/
3. M. Wolpers, M. Memmel, A. Giretti, „Metadata in Architecture Education - First Evaluation Results of the MACE System“, Lecture Notes in Computer Science, vol. 5794/2009, pp. 112–126, Springer Berlin, 2009, ISSN 0302-9743.
4. O. Masutani, H. Iwasaki, „BEIRA: A Geo-semantic Clustering Method for Area Summary“, Lecture Notes in Computer Science, vol. 4831/2007, pp. 111–122, Springer Berlin, 2007, ISSN 0302-9743.
5. Ch. D. Manning, P. Raghavan, H. Schütze, „Introduction to Information Retrieval”, pp. 234–264, Cambridge 2008.
6. „A Wiki for Calls For Papers”. URL: http://www.wikicfp.com/
7. „The DBLP Computer Science Bibliography”. URL: http://dblp.uni-trier.de/
8. „The DBpedia Knowledge Base”. URL: http://dbpedia.org/About/