# Využití struktury webu pro vyhledávání autoritativních institucí a osob

Dalibor Fiala[1], Karel Ježek[1], François Rousselot[2]

[1]Katedra informatiky a výp. techniky, ZČU v Plzni, Univerzitní 22, CZ-30614 Plzeň
dalfia@kiv.zcu.cz, jezek_ka@kiv.zcu.cz

[2]INSA Strasbourg, 24 bd de la Victoire, F-67084 Strasbourg
francois.rousselot@insa-strasbourg.fr

**Abstrakt.** V tomto článku představíme metodologii pro vyhledávání autoritativních vědeckých pracovníků analýzou webových stránek akademických pracovišť. Uvedeme případovou studii zaměřenou na skupinu stránek českých kateder informatiky. Nejprve zanalyzujeme odkazy mezi nimi (jejich vzájemné vztahy) a několika známými hodnoticími algoritmy stanovíme nejvýznamnější katedry. Potom prozkoumáme obsah výzkumných článků nalézajících se na těchto stránkách a určíme nejdůležitější české autory.

**Klíčová slova:** web mining, univerzity, vědci, autority

## 1 Úvod

Pojmy jako důležitost, významnost, autorita, prestiž, kvalita a další synonyma hrají velkou roli ve všech sociálních sítích. Označují objekt, který má velký vliv na ostatní objekty v dané komunitě. Možná nejlepším příkladem jsou bibliografické citace v odborné literatuře. Nyní je tento druh analýzy nezbytný také v oblasti webu. Zde jsou citacemi odkazy mezi webovými stránkami Současné webové vyhledávače proto využívají nejrůznějších hodnoticích algoritmů založených na počítání odkazů, jejichž výsledky kombinují s hledáním slov zadaných v dotazu. Tímto způsobem jsou pak uživateli doporučeny stránky, jež jsou nejen relevantní, ale i vysoce kvalitní. Tyto algoritmy mohou být rekurzivní (např. PageRank nebo HITS [1]) nebo jednoduché (In-Degree). Některé studie [2] prokázaly, že žebříčky „kvality“ generované těmito třemi metodami jsou značně pozitivně korelovány. Nejblíže k naší práci je výzkum prováděný v [4], ale my jsme navíc ke vztahům mezi webovými doménami ještě studovali obsah dokumentů na nich nalezených.

## 2 Experimentální rámec

Naším prvním cílem bylo stanovit autoritativní instituce mezi českými katedrami informatiky. I když jsme naše experimenty omezily obsahem i rozsahem, metodologie, kterou jsme použili, je dostatečně obecná i pro aplikaci v úplně jiném vědním oboru. V prosinci 2005 náš webový pavouk „prolezl“ sedmnáct vybraných serverů. Zajímaly nás pouze odkazy přes protokol HTTP k dokumentům v určitých

formátech. Samozřejmě že počet vstupních odkazů závisí na množství dokumentů na cílovém serveru. Jejich počty se různí v závislosti na odlišných velikostech příslušných institucí, preferenci různých formátů dokumentů a generování dokumentů (dynamické webové stránky) apod. Jeden způsob, jak se vypořádat s touto potíží, je počet citací nějak normalizovat. Např. je možno dělit počet citací počtem dokumentů na citovaném serveru nebo počtem zaměstnanců příslušného pracoviště [4].

## 3 Webové domény autoritativních institucí

Vazby mezi sledovanými servery z tabulky 2 jsou znázorněny na obrázku http://home.zcu.cz/~dalfia/papers/Czech.svg. Citační sítí je orientovaný graf s počty odkazů jako ohodnocením hran. Nejprve jsme spočítali vstupní stupně uzlů v citačním grafu s ohledem (citace) a bez ohledu (in-degree) na hranová ohodnocení. Potom jsme spočítali autority uzlů grafu algoritmem HITS a nakonec vygenerovaly PageRank (přesněji řečeno HostRank) pro každý uzel. Předchozí fáze určování významných institucí nám umožňuje zredukovat množinu webových domén, které budeme analyzovat v následujícím kroku. Například bychom mohli odstranit nejméně důležité servery. Nicméně naše případová studie má dostatečně malá vstupní data, takže žádná redukce není zapotřebí. Měření kvality akademických institucí webometrickými nástroji je zdůvodňováno v [4], kde se žebříčky založené na měření webu velmi podobaly těm oficiálním.

## 4 Autoritativní vědečtí pracovníci podle webu

Kromě zkoumání odkazů ve skupině webových domén kateder informatiky jsme se rovněž zabývali analýzou samotných dokumentů na těchto serverech. Takže mimo soubory obsahující odkazy (především HTML stránky) jsme také stahovali potenciální odborné články. V praxi to znamenalo shromažďovat dokumenty PDF a PS (PostScript), protože většina odborných publikací veřejně dostupných na webu je v těchto dvou formátech. K jejich kategorizaci na články a ostatní jsme použili jednoduché pravidlo. Tímto způsobem jsme nakonec získali asi 3 600 „článků“. Dalším úkolem je extrahovat z článků informace potřebné pro citační analýzu – jména autorů, názvy článků, atd. Zde aplikujeme stejnou metodu jako v [3] založenou na skrytých Markovových modelech (SMM). Z citačního grafu, kde uzly byly „příjmení“ (byly takto označeny klasifikátorem), jsme třemi různými hodnoticími metodami určili nejautoritativnější české autory. Podrobnosti jsou v tabulce 1.

## 5 Analýza vztahu mezi vědci a stránkami

Podle příslušnosti vědců k institucím (zjistíme webovým vyhledávačem) získáme řazení podle autorů. Takže je to žebříček serverů opírající se o citace autorů v publikacích. Tabulka 2 shrnuje všechny žebříčky. Přirozeně nás zajímala míra

**Tabulka 1.** Deset nejvýznamnějších českých autorů. (Jméno může představovat více osob.)

| Pořadí | In-Degree | HITS | PageRank |
|---|---|---|---|
| 1 | Hajič | Kučera | Pokorný |
| 2 | Kučera | Matoušek | Hajič |
| 3 | Nešetřil | Hajič | Jančar |
| 4 | Jančar | Jančar | Matoušek |
| 5 | Matoušek | Nešetřil | Brim |
| 6 | Panevová | Pala | Kučera |
| 7 | Sgall | Smrž | Kratochvíl |
| 8 | Pala | Sgall | Pultr |
| 9 | Kratochvíl | Kratochvíl | Troníček |
| 10 | Smrž | Panevová | Pala |

**Tabulka 2.** Souhrn žebříčků (v závorkách pořadí).

| Doména | Citace | In-Degree | HITS | PageRank | Autoři |
|---|---|---|---|---|---|
| cs.felk.cvut.cz | 6 | 3 | 1 | 4 | 4 |
| iti.mff.cuni.cz | 2 | 6 | 5 | 6 | 6 |
| kam.mff.cuni.cz | 4 | 7 | 8 | 7 | 1 |
| ki.ujep.cz | 14 | 14 | 14 | 14 | 10 |
| kit.vse.cz | 14 | 14 | 14 | 14 | 10 |
| kocour.ms.mff.cuni.cz | 7 | 4 | 4 | 5 | 6 |
| ksvi.mff.cuni.cz | 11 | 9 | 12 | 11 | 10 |
| ktiml.ms.mff.cuni.cz | 13 | 13 | 13 | 12 | 10 |
| ufal.mff.cuni.cz | 14 | 14 | 14 | 14 | 2 |
| www.cs.cas.cz | 8 | 4 | 6 | 2 | 6 |
| www.cs.vsb.cz | 3 | 1 | 2 | 1 | 6 |
| www.fi.muni.cz | 1 | 1 | 3 | 3 | 3 |
| www.fit.vutbr.cz | 9 | 7 | 7 | 8 | 4 |
| www.inf.upol.cz | 11 | 9 | 10 | 9 | 10 |
| www.kai.vslib.cz | 14 | 14 | 14 | 14 | 10 |
| www.kin.vslib.cz | 10 | 9 | 11 | 13 | 10 |
| www.kiv.zcu.cz | 5 | 9 | 9 | 10 | 10 |

**Tabulka 3.** Korelace mezi žebříčky.

| | Citace | In-Degree | HITS | PageRank | Autoři |
|---|---|---|---|---|---|
| Citace | X | 0,89 | 0,89 | 0,86 | 0,63 |
| In-Degree | X | X | 0,96 | 0,96 | 0,65 |
| HITS | X | X | X | 0,95 | 0,64 |
| PageRank | X | X | X | X | 0,63 |
| Authoři | X | X | X | X | X |

korelace mezi danými řadami. Spearmanovy korelační koeficienty pro každý pár se nacházejí v tabulce 3. Všechny jsou významné na 2% hladině. Relativně vysoká korelace mezi řazením podle autorů a ostatními - více než 0,6 - znamená, že velmi citovaní autoři mají pozitivní dopad na významnost webových stránek svých institucí.

## 6 Závěr a další práce

Výsledky, kterých jsme dosáhli, nejsou zcela spolehlivé kvůli omezením a problémům zmiňovaným výše, ale věříme, že naše metodologie je praktická, jak jsme prokázali v našich experimentech. Zdůrazňujeme, že vyvozené závěry jsou zcela závislé na struktuře a obsahu webu, které lze do určité míry snadno ovlivnit, a naším záměrem bylo spíše vyzkoušet navržené obecné postupy v konkrétním prostředí než činit nějaká závažná prohlášení. Uvedený žebříček kateder nebylo možno s ničím srovnat, protože žádný oficiální v ČR neexistuje. Pokud v budoucnu nějaký vznikne, bude zajímavé takové srovnání provést. Taktéž porovnávání žebříčků autorů se chceme věnovat v naší další práci.

Tato práce byla částečně podpořena grantem 2C06009 MŠMT ČR.

## Reference

1. S. Chakrabarti, *Mining the Web: Analysis of Hypertext and Semi Structured Data*. San Francisco, CA: Morgan Kaufmann Publishers, 2003, pp. 209–218.
2. C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, "PageRank, HITS and a Unified Framework for Link Analysis," in *Proc. 25th ACM SIGIR Conf. Research and Development in Information Retrieval*, Tampere, Finland, 2002, pp. 353–354.
3. K. Seymore, A. McCallum, and R. Rosenfeld, "Learning Hidden Markov Model Structure for Information Extraction," in *Proc. AAAI'99 Workshop Machine Learning for Information Extraction*, Orlando, FL, 1999, pp. 37–42.
4. M. Thelwall, "The Relationship Between the WIFs or Inlinks of Computer Science Departments in UK and Their RAE Ratings or Research Productivities in 2001," *Scientometrics*, vol. 57, no. 2, 2003, pp. 239-255.

**Annotation:**

*Using the Web Structure for Finding Authoritative Institutions and People*

In this paper, we present a methodology for finding authoritative researchers by analyzing academic Web sites. We show a case study in which we concentrate on a set of Czech computer science departments' Web sites. We analyze the relations between them via hyperlinks and find the most important ones using several common ranking algorithms. We then examine the contents of the research papers present on these sites and determine the most authoritative Czech authors.