FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2017

Inteligentní vyhledávání dokumentů

Jiří Martínek[1]

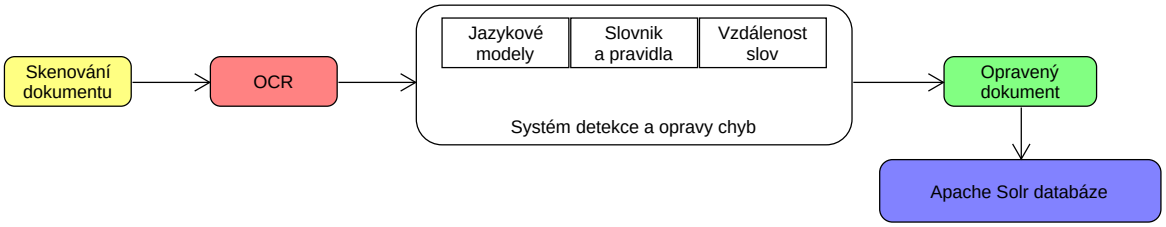
1 Úvod

V současné době je většina dokumentu v nestrukturované podobě. Tato podoba dokumentu je pro počítač nečitelná. Nejčastěji se jedná o naskenované dokumenty, různé ručně psané poznámky či dokumenty staršího data psané na psacím stroji. Tyto dokumenty jsou čitelné pro člověka, ale s jejich zvyšujícím se počtem se zhoršuje schopnost v nich efektivně vyhledávat informace.

Skenované dokumenty je nejprve nutné z obrazové podoby převést do textové pomocí optického rozpoznávání znaků. V rámci převodu bohužel dochází k chybám, proto je nutná existence systému na detekci a korekci těchto chyb. Opravené dokumenty jsou posléze zaindexovány do fulltextové databáze. Vyhledávání je posléze možné prostřednictvím fulltextového dotazu.

2 Hlavní aspekty realizace

Na základě analýz byl jako vyhledávací systém a fulltextová databáze zvolen systém **Apache Solr** (viz Potter (2014)). K realizaci optického rozpoznávání znaků (viz např. publikace Pavlidis (2014)) je použit *open–source* program **Tesseract**. Pro detekci a opravu chyb byla zvolena kombinace strojového učení – **jazykové modelování** (viz Brychcin (2012)) a pravidlového přístupu založeném na slovníku a vzdálenosti slov. Pro zvýšení efektivity vyhledávání a možnosti vyhledávání dle třídy dokumentů byla implementována komponenta klasifikace dokumentů pomocí knihovny strojového učení **Brainy** (viz Konkol (2014)). Souhrnně znázorňuje práci systému obrázek 1.



Obrázek 1: Znázornění práce systému

[1] student navazujícího studijního programu Inženýrská informatika, obor Softwarové inženýrství, e-mail: jimar93@students.zcu.cz

3 Závěr

Oprava chyb přispívá ke zlepšení přesnosti vyhledávání v množině skenovaných dokumentů. Pro ověření úspěšnosti a přesnosti rozpoznávání textu byly vytvořeny experimenty, jejichž účelem bylo otestovat přesnost a chybovost rozpoznávání. V kolekci testovaných obrázků byly dokumenty různých kvalit a obsahů. Experimentálně byla nastavena optimální konfigurace celého systému.

U vstupního textu, kde se objevují cizí jména, názvy, zkratky či celé pasáže v jiném než českém jazyce, vykazuje program horší úspěšnost.

Systém by bylo možné vylepšit přidáním dalších metod strojového učení, například metody sumarizace textu, kdy by byl bezprostředně po fázi rozpoznávání textu vytvořen souhrn, který by obsahoval pouze důležité a klíčové termíny. S takovouto reprezentací by bylo možné snížit velikost indexu a v důsledku toho zefektivnit vyhledávání.

Literatura

Brychcín, T. (2012) *Unsupervised methods for language modeling*: technical report no. DCSE/TR-2012-03

Grainger, T., Potter, T, Seeley, Y. (2014) *Solr in action*

Konkol, M. (2014) *Brainy: A machine learning library*. In: International Conference on Artificial Intelligence and Soft Computing. Springer International Publishing,. p. 490-499.

Pavlidis, T., Mori, S. (1992) *Optical character-recognition*.