Papírové dokumenty v počítači

V ekonomice nebo administrativě stále „obíhá“ **OBROVSKÉ MNOŽSTVÍ PAPÍROVÝCH DOKUMENTŮ**. I proto je převod do digitální formy nesmírně důležitý. Chip vám ukáže, jak celý proces funguje...

PETR KRATOCHVÍL, JÖRG GEIGER

Podzim 1989, Berlín. V centrále východoněmecké tajné služby propuká chaos. Její zaměstnanci mají plné ruce práce při tzv. „systematickém ničení“ choulostivých informací. Výsledek: 16 000 pytlů plných útržků informací o „neoficiálních zaměstnancích“ Stasi, obětech a tajných operacích. Podle odborníků by na ruční sestavení těchto „puzzlů“ lidé potřebovali několik stovek let. Fraunhoferův institut pro výrobní systémy a návrhy technologií (IPK) chce s využitím nejmodernějších technologií dosáhnout stejného cíle do pěti let. Budou použity diagnostické funkce, které se rovněž používají v softwaru pro rozpoznávání textu.

Výzva pro Google

Naučit počítače rozpoznávat znaky se experti pokouší již od poloviny padesátých let. I díky tomu je nyní rozpoznávání textu na velmi vysoké úrovni. Asi nejlepší ukázkou stavu této oblasti je Google. V současnosti se tento vyhledávací gigant „probírá regály knihoven“ a převádí je do elektronické podoby. Jeho cílem je umožnit i vyhledávání v knihách, stejně jako se v současnosti hledá na webu. Další pohled na současný stav vývoje rozpoznávání textu nabízí firma Linguatec, působící v oblasti mobilních technologií. Její software „Shoot & Translate“ nabízí na první pohled neuvěřitelnou věc – pomocí fotoaparátu v mobilním telefonu vyfotíte libovolný cizojazyčný text (například jídelní lístek) a program vám během několika sekund nabídne jeho překlad. Jak to funguje? Software analyzuje snímek, rozpozná v něm text a ten poté přeloží. Technologie rozpoznání textu pomocí optického „skenu“ se označuje jako OCR (Optical Character Recognition).

OCR šetří čas a umožňuje rychlé zadání textu, který byl vytištěn. Byl to právě Lawrence Roberts, který na Massachusetts Institute of Technology (MIT) v Bostonu na počátku šedesátých let vyvinul první procedury pro automatické rozpoznání textu. Základy jeho „technologie“ se využívají dodnes.

Na první pohled to funguje jednoduše. Program porovná načtený znak se všemi uloženými znaky a hledá shodu. V praxi je to ale komplikovanější. Aby byl znak přesně identifikován, musí být shoda jednoznačná. Pak ale stačí, aby byl text kurzivou, a software má problémy. Proto byly vyvinuty speciální fonty (například OCR-A a OCR-B), jejichž jednotlivé znaky abecedy jsou navzájem velmi odlišné. Při jejich použití pak dochází k výborné úrovni rozpoznání. V praxi se ale pracuje s normálními písmy. Teprve v roce 1976 vyvinul Ray Kurzweil první omnifont – tj. OCR systém nezávislý na použitém fontu.

I přes velký pokrok v oblasti OCR nemohou ani dnes skenery „číst“ – bez dodatečného softwaru dokáží přístroje získat pouze „bitmapy“. Do elektronické podoby tak dokáží převést pouze obrázky. Pochopitelně že lze „digitalizovat“ i text, nicméně bez možnosti jeho dalšího použití (například vyhledávání v něm nebo jeho další úpravy). Pro tento účel musí být použity OCR programy, které umí text „číst“ a kontrolovat jeho hodnověrnost. OCR programů existuje opravdu nepřeberné množství – od univerzálních nástrojů pro domácí či kancelářské použití až po profesionální pomocníky nebo specializované utility pro čtení dopisů (ABBYY FineReader OCR XIX) nebo pro rozpoznávání tvarů (například FormPro od OCR Systems).

Rychlík: Scanner knih APT BookScan 2400RA od firmy Kirtas zvládne naskenovat až 1 000 stránek za hodinu.

Jak se rozpoznává text

Automatické rozpoznávání textů pracuje v několika krocích.

KROK 1: URČENÍ ORIENTACE

Staré rčení říká, že s rostoucí kvalitou podkladů roste i kvalita skenování. Aktuální testy ukazují, že znečištěné nebo poškozené podklady mohou velmi znepříjemnit jejich digitální zpracování. Na druhé straně, pokud je kvalita podkladů dobrá, mohou moderní OCR programy zaručit téměř 100% úspěšnost při rozpoznávání textu. Mezi první kroky při zpracování materiálů patří zarovnání stránky – dobré výsledky lze zaručit jen tehdy, pokud je stránka správně vložena do skeneru. Od tohoto bodu se odvíjí další krok OCR programu, který

Tři varianty rozpoznávání textu

Existují tři různé techniky rozpoznávání textu. Moderní „Optical Character Recognition“ (OCR) software obvykle pro zvýšení přesnosti kombinuje všechny…

SEGMENTACE:

Při tomto procesu zjišťuje algoritmus, jak jsou uspořádány oblasti s množstvím barvy a jak oblasti bílé, a tato „zjištění“ jsou statisticky hodnocena. Určitě si rychle všimnete, že barvu najdete uprostřed písmene „A“ a také na jeho vrcholu. Na druhé straně – u písmene „B“ je barva rozdělena především na pravou a levou stranu.

ROZPOZNÁNÍ VZORU:

Při této metodě jsou znaky porovnávány s uloženou sadou možných stylů. Na příkladu vpravo můžete vidět porovnávání naskenovaného písmene s různými styly písmene „A“. Například oblíbený OCR software ABBYY FineReader má ve své databázi pro písmeno „A“ 48 odlišných „šablon“.

ROZPOZNÁNÍ TVARU:

Zatímco u metody „rozpoznání vzoru“ je důležitá shoda s kompletním vzorem, při této metodě se porovnává struktura vzoru, který je rozdělen na jednotlivé části. Například písmeno „A“ tvoří dvě čáry šikmo nahoru a horizontální čára uprostřed. Pokud horizontální čára chybí, pravděpodobně o písmeno „A“ nejde…

kontroluje, zda je stránka vložena správně nebo otočena. Program skáče na různá místa v dokumentu a pokouší se rozpoznat části textu. Pokud není úspěšný, otočí text o 90 stupňů a opakuje předchozí krok. Zde mohou působit problémy dokumenty obsahující zároveň text s horizontální a vertikální orientací. Nakonec software opět prozkoumá několik náhodných vzorků a ověří správnou orientaci textu.

KROK 2: NASTAVENÍ A ZAROVNÁNÍ

Klasické OCR programy vždy pracují s jednou stránkou za druhou. Pokud je zpracována dvojstrana, OCR program rozdělí zdroj na dvě individuální stránky. Problémy se také objevují v případě špatně „přenesených“ faxových dokumentů – poté musí zapracovat korekční algoritmus. Zde je ale nutná mimořádná opatrnost – při neopatrném „čisticím“ a opravném procesu by se „š“ mohlo změnit na „s“ a podobně. Dalším problémem jsou špatně pracující (obvykle ty levné) skenery, které nepracují dostatečně přesně a řádky textů jsou vodorovné jen výjimečně. OCR programy si tedy musí pozici znaků upravovat automaticky.

KROK 3: ANALÝZA USPOŘÁDÁNÍ

Oba předchozí kroky byly jen přípravné, s analýzou uspořádání textu začneme až nyní. Nejprve program provede tzv. segmentaci. To znamená, že určí oblasti stránky, které obsahují text, kde je grafika či tabulky nebo zda se v dokumentu nacházejí speciální prvky, jako čárové kódy či bílé oblasti. Program zkoumá dokument stránku po stránce a u každé z nich rozděluje její strukturu na menší a menší části: od stránky k bloku textu, poté k odstavci, řádku a slovu, nakonec k jednotlivým znakům (viz rámeček nahoře).

Experti nazývají tento proces jako „víceúrovňovou analýzu textu“. Jak ale program rozpozná rozdíl mezi obrázkem a textem? A jak rozpozná odstavce a řádky?

V oblasti rozpoznávání textu došlo v poslední době k celé řadě zlepšení – výsledkem je skutečnost, že programy pracují stále rychleji. V prvním kroku jsou barevné dokumenty převedeny na „černobílé“ (tzv. binarizace). Poté následuje analýza makrostruktury dokumentu.

Tu si lze představit tak, že se na dokument podíváte z větší vzdálenosti přimhouřenýma očima – nemůžete přečíst jednotlivé znaky, ale rozpoznáte strukturu stránky: bloky textu, bílé oblasti a shluky barevných přechodů u obrázků. Ano, právě barevné přechody identifikují obrázky, zatímco proužky symbolů identifikují odstavce textu – tímto způsobem OCR program zjišťuje strukturu stránky. Tento postup pochopitelně může vést i k chybám, především u „lahůdek“ typu „obrázek na pozadí“, ale s tím se programy vypořádají pomocí tzv. vícestupňového přístupu. Hlavní ale je, že se software ze svých chyb učí. Například pokud rozpoznání textu nefunguje korektně a zdroj má i znaky obrázku, software si uvědomí, že s velkou pravděpodobností nejde o text. Prvním krokem

k úspěšnému rozpoznání je tedy identifikace bloků textu. Teprve poté mohou být pomocí mezer oddělena jednotlivá slova a ta rozložena na znaky a body.

KROK 4: ROZPOZNÁNÍ TEXTU

Teprve nyní se dostáváme k jádru věci. V tomto kroku ukáže OCR software své skutečné schopnosti. Pro rozpoznávání znaků existují dvě základní metody: „Pattern matching" (rozpoznání vzoru) a „Feature matching" (rozpoznání tvaru).

Při použití metody „Pattern matching" je znak porovnán se sadou již známých znaků. Pokud znak odpovídá určitému vzoru, je považován za rozpoznaný. Ačkoliv princip této metody zní rozumně, v praxi by samostatně být použita neměla. Důvod: Při aplikaci této metody musí znak odpovídat na 100 procent. Aby tedy bylo rozpoznání „uznáno", musí být k dispozici vzory v příslušných znakových sadách. Ani to však pro úspěšné rozpoznání nestačí – zdroj může být rozmazán, může být napsán kurzivou nebo obsahovat různé velikosti písma. Tyto drobnosti způsobují metodě „Pattern matching" problémy...

Z toho důvodu se v současnosti používá spíše metoda „Feature matching". V ní jsou znaky rozděleny na malé prvky: například malé „b" je rozděleno na svislou čáru a malý půlkruh – to jsou charakteristické vlastnosti tohoto znaku. OCR program tedy zdá všechny „vlastnosti" všech znaků v abecedě.

Nevýhoda: Při použití metody „Feature matching" může být znak rozpoznán pouze s jistou pravděpodobností. To ale není až takový problém, protože OCR software zkouší různé „tvary" a ty se poté snaží zkombinovat „s okolím".

Žádný definitivní algoritmus pro rozpoznávání tedy neexistuje, proto OCR programy kombinují více metod (viz rámeček na straně 57). Některé programy jdou dokonce tak daleko, že používají rozdílné rozpoznávací metody a nechají je pracovat „paralelně". Nakonec program porovná jednotlivé výsledky a rozhodne na základě „většiny".

KROK 5: ZPÁTKY KE ZNAKŮM

Rozpoznání textu je provedeno na úrovni znaků a neméně důležitá je i cesta „zpět" ke slovům. Jako výsledek kroku 4 OCR program „vytvoří" mnoho různých znaků s různou úrovní pravděpodobnosti rozpoznání. Spojíme-li tyto znaky do slov, každé ze slov opět může mít určitou pravděpodobnost. Slova jsou také porovnávána s položkami ve slovníku. V celé řadě programů lze nastavit příslušný jazyk ještě před začátkem skenu, což podstatně zrychluje tento krok. Většina výrobců OCR softwaru nabízí podporu velkého množství jazyků – například u softwaru ABBYY je to téměř 200 jazyků. Nicméně ve „slovnících" se nehledá pouze konkrétní výraz. Hojně se také využívá tzv. morfologických slovníků, které obsahují všechny tvary slov v daném jazyce, a nelze zapomenout ani na porovnání se slovníky „uživatelskými". Posledně jmenované slovníky se využívají například u technických textů – třeba chemici ocení rychlejší rozpoznání slovního spojení „deoxyribonukleová kyselina".

Programy pro rozpoznávání textu však obsahují ještě jeden „kontrolní orgán", který pracuje s pravděpodobností. Ten určuje četnost slova objevujícího se v souvislosti s okolním textem. Například pojem „slavná" památka se v textu objeví pravděpodobněji než „slanná" památka.

KROK 6: FORMÁTOVÁNÍ

Rozpoznávání textu je únavná činnost – neustálé rozčleňování stránek na znaky, jejich detekce a opětovné skládání. Finálním krokem OCR softwaru je nové vytvoření stránky podle původního vzoru. Při této činnosti se využívá znalostí získaných při segmentaci.

Nakonec je určen výstupní formát, kterým může být například doc nebo pdf, přičemž každý OCR program pracuje s několika formáty.

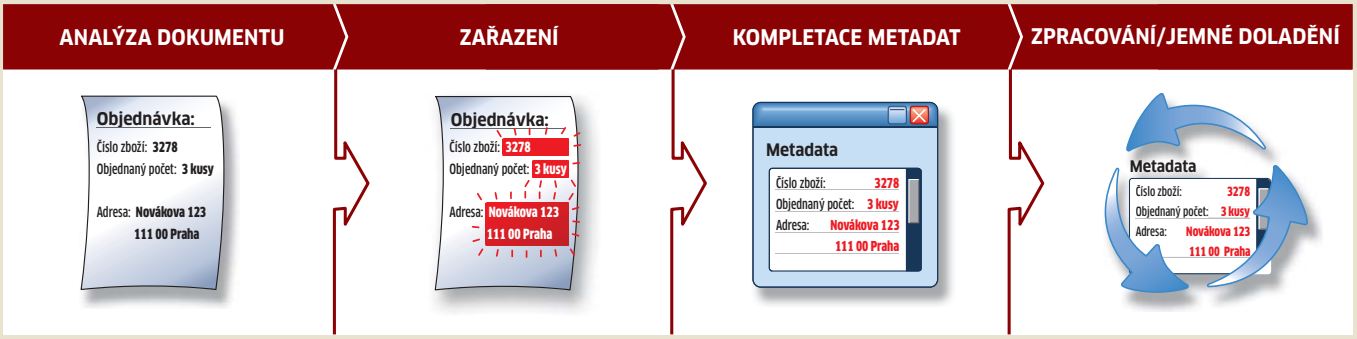
Správa dokumentů

Samotné použití OCR je však obvykle jen polovinou „závodu". Důvod? V dokumentech obvykle bývají chyby, pro pozdější použití je vyžadováno i jejich snadné nalezení – pro tento účel je nutné chytré vyhledávání a kvalitní řízení přístupu.

Problémem ale je, že standardní souborový systém jako NTFS bývá během hledání přetížen, navíc v něm uživatelé mohou hledat jen podle několika málo parametrů – například jména souboru, velikosti nebo data jeho vytvoření. Navíc NTFS a spol. na-

Jak funguje systém správy dokumentů (DMS)

Jedna věc je přečtení dokumentu, porozumění mu je věc druhá: toto schéma vám vysvětlí, jak funguje softwarové rozpoznávání dokumentů.



Nejprve se DMS pokusí načíst dostupná metadata – například informace o autorovi z dokumentu v programu Word. Pokud nejsou žádná metadata k dispozici, jde o prozatím nestrukturovaný dokument.

Systém správy dokumentů prohledává text a hodnotí ho pomocí klíčových slov. To znamená, že slova ve faxu označí jako zprávu.

Nalezené atributy jsou zaznamenány do metabáze. V tomto případě je nestrukturovaný dokument zařazen s dalšími dodatečnými údaji.

Klasifikace však bývá často chybná nebo neúplná. Existují dvě možné cesty vedoucí k opravě metadat: buď pomocí samoučících se systémů, nebo přes ruční opravu.

bízejí jen základní funkce pro řízení přístupu a „operace s verzemi“.

Proto kvalitní programy nabízejí vlastní systémy správy dokumentů (označované jako DMS – Document Management System). Ty umožňují pokročilejší kontrolu přístupu k dokumentům a další pravidla, která se mohou měnit v závislosti na konkrétním uživateli. Jako bonus DMS přidávají vylepšený systém ovládání, stejně jako management metadat a pokročilé indexové vyhledávání. DMS tedy pracuje jako rozhraní mezi uživatelem a „databází“, navíc dokáže plnit další složité úkoly.

Většina současných produktů na trhu v tomto ohledu splňuje normu DFR (Document Filing & Retrieval) ISO 10166, důležitou z hlediska rozsahu a funkčnosti.

A jak vypadá nasazení OCR programů v praxi? Odborníci odhadují, že ve většině firem tvoří 80 % dokumentů tzv. nestrukturovaná data – neboli jde o individuální dokumenty bez vzájemných vazeb a bez možnosti vyhledávání. Příkaz „zobraz všechny dokumenty související se jménem Novák a platbou alespoň 10 000 Kč“ je ve většině firem jen tajným snem – tedy alespoň pokud chcete rychlé a kompletní výsledky.

A právě toto je druhá výhoda DMS databází – mají zvláštní schopnosti dávat řád i zdánlivě nesouvisejícím a nesetříděným datům.

Po vytvoření potřebné struktury lze vyhledávat v mnohem více „informačních polích“, než umožňuje jakýkoliv souborový systém. Lze například vyhledávat podle čísel zákazníků, pořadových čísel nebo osob. Systém zároveň zajišťuje, aby mohli dva lidé zároveň přistupovat ke stejnému dokumentu bez rizika konfliktu. V neposlední řadě také DMS nabízejí vícestupňové archivační systémy, kam se ukládají dokončené a již nepoužívané dokumenty.

Struktura DMS dokumentů

Ve srovnání s klasickými databázemi nabízejí DMS celou řadu výhod. Již zmiňovanou specialitou je možnost dát strukturu neorganizovaným dokumentům.

Příklad: Firma obdrží objednávku faxem. Tento dokument je naskenován a prostřednictvím OCR načten do DMS. Až do tohoto bodu je dokument nestrukturovaný. DMS poté spustí prohledávání dokumentu na klíčová slova. Zde jsou použity tzv. klasifikační systémy, které dokáží rozlišit faktury od objednávek. Pochopitelně jsou možné i manuální zásahy do DMS, protože tyto systémy poskytují pouze „pravděpodobné“ výsledky, a ty nemusí být stoprocentní. Systémy, které pracují s vektory podobnosti nebo neuronovými sítěmi, jsou ještě o krok dále a za provozu se stále učí. Nicméně největší výhodou pro uživatele je stále možnost „univerzálního prohledávání“. V celé řadě firem už je nemyslitelné, aby zaměstnanci nemohli zadávat dotazy pomocí klíčových slov „množství“, „faktura číslo“ nebo „jméno zákazníka“. DMS shromažďuje všechny informace v samostatném souboru pro metadata. V současnosti jsou namapovány v XML na základě návrhu „Definice typu dokumentu“. Kromě metadat DMS připojuje i tzv. vyhledávací index, který lze přirovnat k indexu knihy v knihovně (viz rámeček dole). Při vyhledávání pak DMS místo prohledávání celé databáze analyzuje index a správnou odpověď tak uživateli nabídne rychlostí blesku.

Trendy, výzvy a poklady

V současnosti patří mezi výzvy v oblasti rozpoznávání textu především velké projekty typu Google Books nebo digitalizace knih ve velkých knihovnách. Česká republika patří v podobných projektech mezi špičku – například Národní knihovna České republiky vede celoevropský digitalizační projekt ENRICH. V pražském Klementinu existuje špičkové digitalizační pracoviště, jehož skenery již prošly klenoty související s historií českého národa. V digitální podobě je tak již například Vyšehradský kodex, Kosmova kronika nebo známá „Ďáblova bible“. Podobné digitalizační aktivity vyvíjí i Městská knihovna v Praze, která se však specializuje na beletrii a dokumenty související s hlavním městem. Za perlu a unikát v oblasti využití OCR lze označit projekt Gutenberg, který v roce 1971 odstartoval Michael Hart, student univerzity v Illinois. Od té doby on a tisíce dalších dobrovolníků zdigitalizovalo téměř 30 tisíc knih, které jsou na webu projektu (**www.gutenberg.org**) zdarma k dispozici.

Za více než půl století se digitalizace dostala na výbornou úroveň, přesto však lze nalézt oblasti, kde je ještě co zlepšovat. OCR programy mají například stále ještě problémy u starších tisků a výrobci softwaru stále pracují na lepším paralelním rozpoznávání. I tak je ale cesta od klasického papíru k nulám a jedničkám snadnější než kdykoliv předtím.

AUTOR@CHIP.CZ

Jak se vytváří vyhledávací index

Indexovací soubor je vytvářen i s pomocí počítání a třídění. Díky tomu může být později nalezena požadovaná informace velmi rychle.

Indexovací služba „systému správy dokumentů“ nebo služba Windows „na pozadí“ prohledává celý dokument.

Index je nastaven paralelně k dokumentu, tak jako obsah v případě knihy. V tomto souboru je také uvedeno umístění jednotlivých slov.

Toto zobrazení pomocí indexu přesně odpovídá obsahu dokumentu. Nicméně komprimovaný soubor je mnohem menší než samotný text.

INTERNETOVÉ ODKAZY:

www.prahavknihovne.cz
Více než 1 000 zdigitalizovaných knih souvisejících s Prahou.

www.ipk.fraunhofer.de/en
Informace o puzzlech z dokumentů tajné služby Stasi a metodách využití OCR.

www.linguatec.net
Nabídka softwaru pro mobilní telefony využívající OCR pro překlady.

http://recaptcha.net/
Metoda ochrany proti spamu – captcha – pomáhá s digitalizací starých knih.

www.prahavknihovne.cz/Jak-se-digitalizuje-kniha-16.htm
Informace a videoreportáže z oblasti digitalizace.