ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS
SBORNÍK MENDELOVY ZEMĚDĚLSKÉ A LESNICKÉ UNIVERZITY V BRNĚ

Ročník LVII 32 Číslo 6, 2009

STANOVENÍ METOD AUTOMATIZOVANÉHO HODNOCENÍ FORMÁLNÍ KVALITY DOKUMENTŮ

P. Talandová, J. Rybička

Došlo: 11. května 2009

Abstract

TALANDOVÁ, P., RYBIČKA, J.: *Method specification for automated evaluation of documents formal quality.* Acta univ. agric. et silvic. Mendel. Brun., 2009, LVII, No. 6, pp. 305–314

Automated documents processing allows production of large amount of documents. Formal quality of the documents is very important as it contributes to better understanding and information transmission. The paper deals with the automated documents quality evaluation. This requires a design of a document model. The model contains the objects of which the pages are compiled, the types of objects and, the most important, the objects' parameters. The parameters of the object are very important as they are inputs for the document evaluation according to the typographical rules. The parameters are an important part of the model which should reliably describe the document. A set of criteria is designed, which are used to describe the requirements on appropriate methods for model formation. From large amount of methods, methods that meet the criteria can be applied to the document. The result is a model of a real document which can be used for the automatic evaluation based on the typographical rules.

document quality, formal quality, automated evaluation, document model, physical analysis, logical analysis, typography

V souvislosti s počítačovým zpracováním textů je často diskutovaným pojmem „kvalita dokumentů". Důraz je přitom kladen nejen na kvalitu obsahovou, ale zejména na kvalitu z hlediska struktury, formální úpravy a typografie. Vzhledem k tomu, že autorům nejrůznějších obchodních sdělení, marketingových materiálů, odborných a vědeckých článků či výukových textů již nic nebrání publikovat dokumenty v elektronické podobě bez další kontroly zodpovědnou osobou, je nanejvýš potřebná kontrola těchto dokumentů z hlediska jejich formální úpravy. Formální aspekty přispívají k lepší srozumitelnosti textu, a tedy snazšímu předání informací, proto je nezbytné jejich kontrole a hodnocení věnovat zvýšenou pozornost. Speciální význam má toto hledisko u odborných a vědeckých prací, například závěrečných kvalifikačních prací (bakalářských, diplomových), kde je formální kvalita nezbytnou nutností a také jedním z kritérií při konečném hodnocení práce (Haluza a kol., 2008).

Kvalita formální stránky dokumentů je zcela samozřejmým požadavkem také v případě textové komunikace v podnikové praxi (nejčastěji v podobě nejrůznějších marketingových materiálů). Požadavek na automatizaci procesu hodnocení formální kvality zde pramení jednak z relativně velkého objemu takových dokumentů, ale také z absence všeobecného povědomí tvůrců (autorů) o pravidlech, zásadách a technologiích zajišťujících dostatečnou formální úroveň.

Automatizace hodnocení formální kvality dokumentů vyžaduje jejich strojovou analýzu. Určitá analýza dokumentů je sice běžně prováděna (podrobněji v Talandová, 2007), avšak ne vždy je zaměřena komplexně; obvykle se soustředí pouze na některé aspekty dokumentu (např. systém OCR nebo analýza použitého písma). Pro analýzu stránky a dokumentu je však třeba vybrat metody, jejichž výsledky budou vhodným vstupem pro hodnocení kvality sazby. Cílem tohoto článku je návrh modelu dokumentu vhodného pro hodnocení formální

stránky a na to navazující stanovení metod analýzy dokumentů z hlediska jejich přínosu a využitelnosti při hodnocení formální kvality.

METODIKA

Pro hodnocení kvality dokumentu je nezbytné nejprve definovat obsah pojmu *kvalita dokumentu* a rozhodnout, jak bude dokument pro účely hodnocení kvality popsán.

Vzhledem k automatizovanému zpracování je třeba sestavit formální *model dokumentu D*. Dokument je podroben analýze, při níž se identifikuje množina prvků stránek *P*, u nichž je určen typ prvku z množiny *T* a jeho parametry (množina A_r).

Pro analýzu dokumentů existuje řada metod a stále přibývají další. Pro získání požadovaných informací o dokumentu je tedy podstatný *výběr vhodných metod* analýzy, přičemž za vhodné jsou považovány ty metody, které splňují zadaná kritéria. Na použitých metodách závisí to, jaké informace je možné z dokumentu získat. Množina dostupných metod $W = \{w_1, \ldots, w_x\}$ tedy podléhá hodnocení, jehož cílem je nalezení takové podmnožiny metod $W_1 \subset W$, které poskytují výsledky v souladu se zvolenými kritérii z množiny *K* (Obr. 1).

V tomto článku je navržena metodika pro výběr a hodnocení vhodnosti metod analýzy. Podle metodiky budou následně metody posuzovány a metody z podmnožiny vyhovujících metod W_1 budou moci být použity pro provedení vlastní analýzy dokumentu, jejímž výstupem jsou informace o kvalitě.

Kvalita dokumentů

Při hodnocení *kvality dokumentů* je třeba specifikovat význam pojmu *kvalita* a metody analýzy vybírat s ohledem na obsah tohoto pojmu. Pozornost je přitom soustředěna výhradně na formální kvalitu, tj. na úpravu, vzhled dokumentu a jeho typografické zpracování, nikoli na kvalitu obsahu.

Pojem *dokument* se vztahuje k tištěným či elektronickým dokumentům, které obsahují převážně text a jejichž obsah je uspořádaný do stránek. Výklad pojmu *kvalita* uvádí norma ČSN EN ISO 9000. Podle definice jde o *stupeň splnění požadavků souborem inherentních charakteristik*.

Na základě těchto definic lze kvalitu dokumentu vymezit jako *stupeň shody daného dokumentu s pravidly pro formální úpravu*.

Pravidla pro úpravu jsou dána dlouholetými zvyklostmi a odrážejí také fyziologii čtenáře. Vztahují se k různým částem (úrovním) dokumentu, které jsou zohledněny v modelu dokumentu.

Vybrané vlastnosti kvalitního dokumentu

Kvalita dokumentu je dána mírou shody vyjádření (reprezentace, zobrazení) prvků dokumentu s požadovanými a obecně uznávanými pravidly pro jejich vyjadřování. Požadavky se zaměřují na klíčové vlastnosti dokumentů, jejichž porušením by dokument trpěl nejvíce. Vlastnosti, které se nejčastěji vyskytují v odborné literatuře a jejichž důležitost je ověřena i vlastními poznatky, jsou dále stručně popsány.

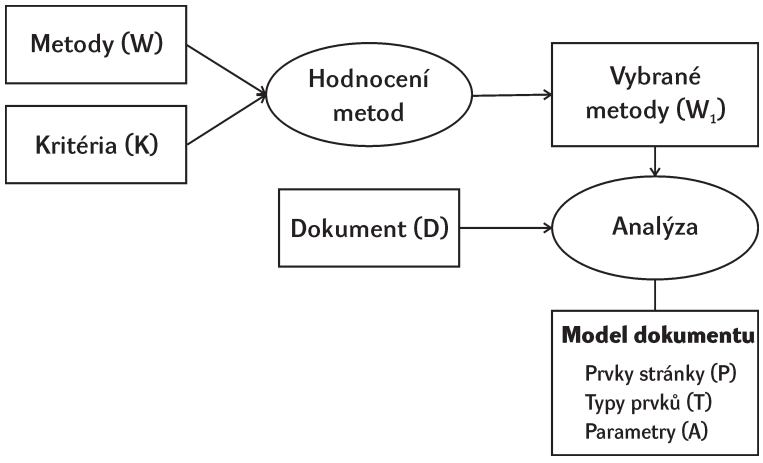
Symetrie a zarovnání souvisí s estetickým uspořádáním stránky. Dle zkušeností působí symetrická stránka se zarovnanými objekty na čtenáře lépe a je čitelnější, proto je věnována pozornost rozmístění prvků a jejich uspořádání. Dokument s kvalitní sazbou by se měl vyznačovat velkou mírou symetrie a zarovnaných objektů (Doermann a kol., 1998).

Rovnoměrnost a vyváženost zahrnují rovnoměrné pokrytí plochy stránky jednotlivými prvky, uspořádání a umístění prvků v ploše.

Oddělení a seskupení prvků zahrnuje práci s nepotištěným místem na stránce. I tato nepotištěná, tzv. bílá místa mají svoji funkci, vizuálně naznačují logickou vazbu mezi prvky a jejich souvislosti (Harrington a kol., 2004) a usnadňují orientaci čtenáře v textu.

Proporce prvků souvisejí s estetikou dokumentu a s jeho čitelností. Jde zejména o poměr délek stran objektů i stránky a umístění nejdůležitějších objektů do optického středu stránky.

Grafické hledisko zahrnuje práci s barvami i s grafickými objekty a opět souvisí s estetickou úrovní dokumentu (Felici, 2003). Na barvy je kladen důraz zejména v souvislosti s čitelností.



1: *Hodnocení metod a jejich využití pro analýzu dokumentu*

Tyto vlastnosti jsou pro kvalitní dokument nezbytné a směřují k naplnění požadavku *čitelnosti*, což je jedna z nejdůležitějších charakteristik dokumentu.

Požadavky na kvalitu jsou vymezeny uvedenými vlastnostmi kvalitního dokumentu. Jedná se však o poměrně abstraktní pojmy, které je třeba u jednotlivých prvků dokumentu konkretizovat v podobě parametrů a jejich hodnot. Tato konkretizace je důležitá, protože s nevhodně zvolenými, zkreslenými či chybějícími parametry by informace o kvalitě byly irelevantní.

Dokument včetně typografických pravidel je tedy nezbytné formalizovat – model dokumentu a model typografických pravidel představuje materiál umožňující automatizované porovnání, a tím i určení míry shody, tj. kvality dokumentu.

VÝSLEDKY

Formální model dokumentu

Dokument lze chápat jako strukturu, která je sestavena z jednotlivých stránek. Každá stránka obsahuje prvky různých typů. Dokument jako celek, jeho stránky a prvky stránek lze vyjádřit stromovou strukturou. Kvalita dokumentu je ovlivněna hodnotami parametrů na úrovni prvků stránky, jednotlivých stránek i dokumentu.

Nejnižší úroveň dokumentu, která byla pro účely této práce vymezena, představují *prvky na stránce*. Každý prvek náleží k určitému typu a je popsán množinou parametrů (atributů) závisejících na tomto typu. Parametry (resp. jejich konkrétní hodnoty) mají přímý vliv na kvalitu daného prvku.

Následující úrovní, na níž se projevuje kvalita zpracování dokumentu, je úroveň *stránky* a s tím související stránková sazba. Kvalita na této úrovni je ovlivněna kvalitou jednotlivých prvků, nejedná se však o prostý souhrn. Významný je také způsob uspořádání prvků v rámci stránky, je třeba řešit vztah prvků navzájem i jejich vztah ke stránce jako nadřízenému prvku.

Nejvyšší úrovní je samotný *dokument*, který je představován posloupností stránek. Také v tomto případě platí, že souhrn kvalitních stránek nestačí pro vznik kvalitního dokumentu, vliv má i jejich vztah a vzájemné působení. Teprve v případě, že jsou splněna pravidla pro kvalitu na předcházejících úrovních i na úrovni dokumentu, lze hovořit o kvalitním dokumentu.

Každá z těchto úrovní vyžaduje jiný přístup, proto je nezbytné rozdělit způsoby analýzy a zjišťování kvality s ohledem na jednotlivé úrovně.

Prvky na stránce a jejich typy

Pro podrobné posouzení kvality je třeba provést analýzu dokumentu a stránky a rozlišit prvky stránky. V prvé fázi hodnocení je rozpoznávána pouze struktura stránky, provádí se tzv. *fyzická analýza*. Prvky jsou vyhledávány podle svých fyzických vlastností, zejména podle rozměrů a umístění. Na stránce P je identifikován seznam m prvků p_1 až p_m. Stránku lze tedy chápat jako množinu $P = \{p_1, \ldots, p_m\}$.

V další fázi se určuje typ, význam a funkce těchto prvků, provádí se tzv. *logická analýza*. Nechť existuje množina $T = \{t_1, \ldots, t_n\}$, kde n je počet typů prvků a t_1 až t_n představují konkrétní typy prvků. Každý z typů t_1 až t_n je definován svým názvem a množinou vlastností, které identifikují daný typ a odlišují jej od ostatních typů.

Typická množina T obsahuje následující typy prvků: odstavce všech druhů (běžný text, citát, výpis, apod.), nadpisy všech úrovní, seznamy (číslované, nečíslované, popisné), matematické, chemické a fyzikální výrazy a vzorce, tabulky, obrázky a další grafické prvky, popisky (pro tabulky nebo obrázky), poznámky pod čarou, marginálie, záhlaví a paty stránky.

Pro každý druh dokumentu lze konstruovat odpovídající množinu T, přičemž prvky této množiny jsou vybrány před zahájením logické analýzy v závislosti na znalosti druhu dokumentu.

Cílem logické analýzy je evaluace prvků p_1 až p_m, tj. zobrazení e: $P \rightarrow T$. Platí, že všechny prvky množiny T nemusejí být na stránce obsaženy, ale každý prvek stránky musí být svázán s právě jedním typem.

Dokument jako celek lze interpretovat jako množinu stránek $D = \{P_1, \ldots, P_s\}$ a každá stránka obsahuje m_i prvků, kde i je prvek z $\{1, \ldots, s\}$. Pak jednotlivé stránkové prvky v rámci dokumentu tvoří množinu

$$P_D = \bigcup_{s=1}^{s} P_i = \{p_{11}, p_{12}, \ldots, p_{1m_1}, p_{21}, \ldots, p_{sm_s}\}$$

Parametry prvků dokumentu

Parametry vycházejí především z používaných typografických pravidel a jsou navrženy tak, aby vystihovaly charakter jednotlivých typů prvků. Část parametrů je společná pro všechny typy prvků, kromě toho má každý typ prvku také další parametry.

U prvků všech typů se evidují tyto parametry:

- rozměry – výška a šířka,
- nepravidelnost okrajů (parametr, který ovlivňuje rozměry a zarovnání prvků),
- relativní charakteristiky prvku (poměr stran, plocha prvku, poměr plochy prvku vůči ploše stránky),
- intenzita zabarvení,
- umístění na stránce (pro tyto účely bude zaveden referenční bod prvku, který bude totožný se souřadnicemi levého horního rohu prvku).

Další parametry závisejí na typu prvku.

Pro *textové typy* prvků (odstavce, nadpisy, seznamy, popisky, poznámky pod čarou a marginálie) se evidují i parametry odstavcové sazby. Určují se především:

- parametry písma (rodina písma, řez, stupeň, barva),
- intenzita zabarvení textu na detailnější úrovni a velikost mezer,

- údaje specifické pro textové prvky (rozměrové parametry odstavce, např. zarážka).

U *seznamů* se eviduje typ číslování nebo jiného označení položek seznamu a úroveň seznamu.

U *tabulek* je důležité především množství světla v tabulce. Kromě již zmíněných parametrů je vhodné sledovat sílu linek.

U *obrázků* a *grafických prvků* lze evidovat informace o barvách, popř. o tvaru.

Uvedené vlastnosti tvoří množinu parametrů A, přičemž množina je rozšiřitelná o další vlastnosti. Pro každý typ prvku $t_r \in T$, kde $r \in \{1, \ldots, n\}$, lze konstruovat množinu A_r s parametry a_1 až a_c (c je celkový počet parametrů). Pro každý prvek p_{smr} existuje množina A_{smr} s konkrétními hodnotami parametrů.

Stránka jako prvek dokumentu

Zvláštním případem při hodnocení dokumentu je stránka dokumentu, která se dá při jisté úrovni abstrakce považovat za jeden prvek. Stránka bude posuzována stejně jako jiné prvky.

Stránka je popsána pomocí parametrů, které zahrnují:

- výšku a šířku, poměr stran a plochu stránky,
- intenzitu zabarvení,
- rovnoměrnost a symetrii (pokrytí plochy stránky prvky),
- rozměry okrajů.

Dále je vhodné evidovat seznam prvků na stránce, jejich typy a uspořádání.

Metody pro zjištění parametrů stránky jako jednoho komplexního prvku jsou obdobné jako metody pro získání parametrů jiných prvků. Uvedené parametry stránky lze dále využít při hodnocení vztahů jednotlivých prvků ke stránce.

Formalizace typografických pravidel

V uvedeném modelu se jeví jako základní element parametr stránkového prvku $a_i \in A_{smr}$, kde $i \in \{1, \ldots, c\}$. Modelem konkrétních typografických pravidel je *rozsah povolených hodnot* daného parametru. Vyhodnocení formální přijatelnosti daného prvku spočívá v určení míry shody získaných konkrétních parametrů prvku dokumentu s typograficky podloženými a stanovenými rozsahy.

Dalším specifickým typografickým požadavkem je jednotnost prvků.

Jednotnost prvků

Požadavky na kvalitu stránkové a odstavcové sazby zastřešuje požadavek *jednotnosti*. Prvky téhož typu by měly mít stejné vlastnosti, jednotnost má být dodržena i mezi prvky různých typů. Zpracování na úrovni prvků, v rámci stránky i v rámci celého dokumentu má být konzistentní (Felici, 2003). Tuto skutečnost lze v našem modelu stránkových prvků vyjádřit tak, že je-li $e(p_{ij}) = e(p_{kl})$, pak pro libovolné i, j, k, l (i a k jsou z $\{1 \ldots s\}$, j a l jsou z příslušných $\{1 \ldots m_j\}$ a $\{1 \ldots m_l\}$) jsou prvky p_{ij} a p_{kl} zobrazovány identicky. Tuto identitu je potřebné chápat na úrovni typu, tj. nejde o identický obraz, ale o identický způsob zobrazování, tedy o identický soubor parametrů.

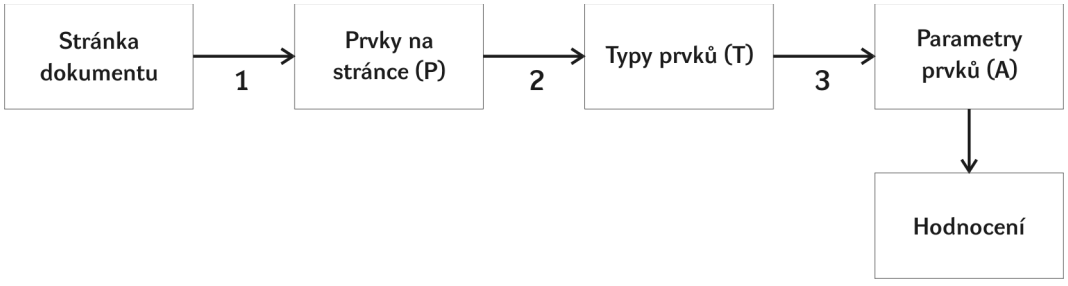
Požadavek jednotnosti je jedním z nejdůležitějších pravidel pro úpravu dokumentů. Míra identického zobrazení prvků má přímou souvislost s kvalitou dokumentu.

Odvození metod pro sestavení modelu reálného dokumentu

Uvedené pohledy na jednotlivé úrovně a jejich vlastnosti se promítají do sestavení modelu reálného dokumentu. Modelový pohled na dokument zahrnuje prvky na stránce, jejich typy a podle typů také příslušné parametry. Tyto dílčí charakteristiky, stejně jako komplexní vlastnosti stránky, jsou dále předmětem hodnocení kvality.

Sestavování modelu zahrnuje tři fáze (Obr. 2).

1. Ve *fázi 1* je třeba provést fyzickou analýzu, jejímž cílem je segmentace, nalezení prvků na stránce. Tato oblast je poměrně propracovaná a existuje řada postupů, které je možné pro tyto účely použít. Převážně se jedná o metody analýzy obrazu, jejichž cílem je rozlišit objekty od nepotištěných míst na stránce (Cattoni, 1998; Harrington, 2004; Mao a kol., 2003; Slocombe a Ambekar, 1998). Lze však i využít jiných postupů a zkoumat vstup ve formátu XML (Fuss a kol.; Hitz, 1999) nebo PDF (Chao a Fan, 2004; Lovegrove a Brailsford, 1995; Rigamonti a kol., 2005). Výsledkem je popis stránky obsahující jednotlivé prvky (množina P).
2. *Fáze 2* představuje logickou analýzu, tj. nalezení zobrazení e: $P \rightarrow T$. Je třeba provést klasifikaci a rozdělit prvky stránky do stanovených tříd podle typů, které tvoří množinu T. Tato klasifikace je podstatná pro další zpracování, přede-



2: *Schéma sestavení modelu dokumentu a přechod od stránky k jejímu hodnocení*

vším pro získání parametrů prvku. Metody, které mohou být pro tento krok analýzy použity, proto musejí odpovídat stanoveným kritériím.

3. Ve *fázi 3* je třeba pro každý prvek p_{smr} získat množinu A_{smr} s konkrétními hodnotami stanovených parametrů. Parametry prvků jsou vstupem pro hodnocení kvality dokumentu, proto je nezbytné, aby hodnoty parametrů byly zjištěny co možná nejpřesněji a nejúplněji. Pro získání hodnot parametrů jednoho prvku se proto využívá více než jedné metody. I v tomto případě lze využít jen metody, které odpovídají stanoveným kritériím.

Výběr kritérií pro hodnocení metod

Pro určení typu prvku a pro získání hodnot parametrů prvku lze použít ty metody, které vyhovují *kritériím pro hodnocení metod*. Kritéria jsou volena tak, aby bylo možné posoudit, zda daná metoda může přispět k sestavení *modelu stránky*, který je výchozím bodem pro další hodnocení. Kritéria vycházejí z požadavků na vytvářený model. Jedná se o minimální sadu kritérií, která musí splňovat každá použitá metoda, v souvislosti s dalším vývojem lze přidávat nová kritéria.

Některá kritéria se týkají pouze klasifikace prvků nebo pouze získání hodnot parametrů, část kritérií je společná. Označme $K_T = \{k_1, \ldots, k_f\}$ množinu kritérií, která se týkají metod pro klasifikaci prvků, a $K_A = \{k_1, \ldots, k_g\}$ množinu kritérií, která se týkají metod pro získání hodnot parametrů.

Základní navržená kritéria uvádí následující výčet.

- *Automatizovatelnost*. Metoda musí být plně automatizovatelná, aby bylo možné zjistit požadované informace pouze na základě analýzy vstupu, bez zásahu uživatele. Při výběru metod je automatizovatelnost nutnou podmínkou.
- *Efektivnost*. Metoda pracuje efektivně, s přiměřenou prostorovou i časovou náročností. Upřednostňují se ty metody, které jsou schopny získat z dokumentu více informací.
- *Práce s neúplnými a nepřesnými daty*. Metoda je schopná pracovat i tehdy, pokud se některé informace nepodaří zjistit nebo se je podaří získat pouze přibližně. Metoda umí zobecňovat a odhadovat (doplňovat) neúplná data a pracovat s nejistými informacemi. Dokáže nalézt dostatečně přesné řešení problému.

Kritéria pro klasifikaci prvků:

- *Spolehlivost*. Přiřazení typu k prvku na stránce je dostatečně spolehlivé, typ prvku je určen správně. V opačném případě by další hodnocení neodpovídalo realitě a zkreslilo by obraz dokumentu.
- *Využívání vzorů*. Metoda umí využít podobnosti prvků, prvky obdobných charakteristik klasifikuje stejně.

Kritéria pro získání hodnot parametrů:

- *Spolupráce metod*. Metoda dokáže spolupracovat s jinou metodou – alespoň v tom smyslu, že se metody mohou podílet na získání hodnot parametrů pro jeden prvek.
- *Typ prvku*. Metoda respektuje typ prvku a dokáže získat hodnoty parametrů příslušejících k tomuto typu prvku.
- *Úplnost*. Existuje kombinace metod, která pro každý prvek p_{smr} zjistí hodnoty jeho parametrů $a_c \in A_{smr}$.

Metody pro analýzu dokumentů

Soubor metod používaných pro analýzu dokumentů je poměrně rozsáhlý. Metody lze kategorizovat z několika hledisek (např. podle formátu vstupních dat), přičemž v této práci je pozornost soustředěna na *metody pro identifikaci typů prvků stránky* (množina W_T) a *metody pro získání parametrů prvků* (množina W_A). Platí, že $W_T \cup W_A = W$.

Metody pro identifikaci typů prvků stránky

Pro zjištění typu prvku je vyžadováno, aby metoda $w_x \in W_T$ splňovala následující kritéria K_T: automatizovatelnost, efektivnost, spolehlivost, přínos pro vytvoření modelu stránky, práci s neúplnými a nepřesnými daty a využívání vzorů.

Pro tento typ úlohy je vhodná skupina metod, které dokážou vstupní data podle daných požadavků klasifikovat. Pro klasifikaci je určena množina nepřekrývajících se kategorií (zde typů) a prvek je zařazen do jedné z těchto kategorií. Dále je požadováno, aby metoda uměla využívat vzory a obdobné prvky zařazovala do stejné skupiny. V ideálním případě se zařazování zpřesňuje a prvky jsou klasifikovány s větší úspěšností.

Tento druh úlohy je vhodný pro zpracování *neuronovou sítí* (Gori a kol. 2003; Mařík, 1993). Neuronová síť zároveň splňuje požadavky kladené na metody. Lze uvažovat i o zavedení expertního systému, jedná se však již o komplexní řešení.

Metody pro získání parametrů prvků

Po určení typu prvku lze přistoupit k získání parametrů prvku. Tato fáze je závislá na správném určení typu prvku, od něhož se odvíjí množina parametrů pro každý prvek.

Pro zjištění parametrů je vyžadováno, aby metody splňovaly následující kritéria K_A: automatizovatelnost, efektivnost, předpoklady pro práci s nepřesnými daty, přínos pro vytvoření modelu stránky a především respektování typu prvku a schopnost zjistit jeho jednotlivé parametry. Dále je třeba zajistit, aby bylo možné kombinovat jednotlivé metody a tím získat hodnoty všech parametrů.

Prvky i různých typů mají některé parametry stejné. Jde především o rozměry a umístění prvku na stránce, intenzitu zabarvení a pravidelnost okrajů. Pro zjištění těchto parametrů lze využít informace získané v předchozích fázích rozpoznávání a dále skupinu metod pro analýzu obrazu (např. Slocombe a Ambekar, 1998; van Beusekom, 2006).

U textových typů prvků se dále sledují charakteristiky na úrovni odstavcové sazby (Bapst a Ingold,

1998; Ma a Doermann, 2005; Richy a André, 1996; Sennhauser, 1993). Pro zjištění těchto informací se používají metody analýzy obrazu, nebo se využívá informací uvedených přímo v dokumentu.

Zmíněné metody splňují požadavek automatizovatelnosti. Významně pomáhá znalost typu prvku tím, že přesně vymezuje parametry, které je třeba zjistit. Toto řešení dává předpoklady pro efektivní fungování metody. Při výběru metod pak zbývá posoudit, nakolik metoda umí pracovat i s nepřesnými daty, a jaké výsledky metoda přináší ve srovnání s jinými metodami. Bude upřednostňována taková kombinace metod, která bude přinášet co nejlepší výsledky při co nejmenší náročnosti zjišťování.

DISKUSE

Pro hodnocení dokumentů je k dispozici množina metod $W = W_T \cup W_A$ a množina kritérií $K = K_T \cup K_A$. Výstupem je množina vhodných metod $W_1 = W_{1T} \cup W_{1A}$. Navržená metodika stanovuje, jak posuzovat dostupné metody podle vybraných kritérií. Návrh platí pro metody a kritéria jak ve fázi určování typů, tak ve fázi získávání parametrů.

Aby mohla být metoda použita pro práci s dokumentem, musí splňovat všechna požadovaná kritéria alespoň do určité míry. Kritériím lze přiřadit pravdivostní ohodnocení z množiny {0;1} nebo z intervalu <0;1>, kde 0 znamená „nesplněno“ a 1 znamená „zcela splněno“. U všech kritérií z množiny K lze doporučit hodnocení v rozsahu <0;1>, kde platí přímá úměrnost mezi mírou splnění kritéria a jeho číselným ohodnocením. Výjimkou může být kritérium, které požaduje respektování typu prvku a zjišťování parametrů příslušejících k prvku. Je-li to možné, měly by být vyloučeny metody, které u posuzovaného prvku nezjistí hodnoty jemu příslušných parametrů.

Pro každou metodu $w_x \in W$ bude vytvořeno ohodnocení kritérií z množiny K. Toto ohodnocení nelze provést automatizovaně. Ohodnocení by měl provádět expert, který je obeznámen s metodou a jejími principy. Každá nová metoda musí být před zařazením do množiny W takto ohodnocena.

Ohodnocení hlavních kritérií, zahrnutých do množiny K, je možné specifikovat pomocí dílčích (podřízených) kritérií, jejichž hodnoty ovlivňují hodnotu nadřízeného kritéria. Tento vztah je rekurzivní. Existuje tedy strom kritérií, přičemž koncové uzly (listy) jsou tvořeny elementárními kritérii s pravdivostním ohodnocením z množiny {0;1}. V mezilehlých uzlech stromu se tato kritéria kombinují podle zvolené funkce (např. lze počítat průměr hodnot v bezprostředně podřízených uzlech stromu). Tímto způsobem je určena i hodnota hlavních kritérií $k_f \in K_T$, resp. $k_g \in K_A$, která se využívají pro další hodnocení. Další aplikací téže funkce lze získat číselné ohodnocení, které charakterizuje metodu jako celek.

V praxi mohou mít kritéria různou důležitost, která bude při hodnocení metod realizována pomocí nastavení vah. Váhy budou stanoveny pro každý případ hodnocení metod jednotně a budou přiděleny všem hodnoceným metodám stejně, aby byla zajištěna srovnatelnost hodnocení.

Váhy tvoří vektor vah $V_T = \{v_1, \ldots, v_f\}$, resp. $V_A = \{v_1, \ldots, v_g\}$. Hodnoty jednotlivých kritérií k budou vynásobeny hodnotami příslušných prvků vektoru V. Získané výsledky lze použít k výběru metod, které splňují všechna kritéria (množina W_1).

Pro metody $w_x \in W_T$ platí, že metoda bude zařazena do množiny W_{1T}, pokud splňuje všechna zvolená kritéria. Kritérium považujeme za splněné, pokud jeho hodnota přesahuje zvolenou prahovou hodnotu z ($z > 0$).

$$\forall\, w_x \in W_T\colon w_x \in W_{1T} \leftrightarrow \forall\, k_f \in K\colon k_f > z$$

Je-li více takových metod, bude vybrána ta, která dosahuje celkových lepších výsledků, popř. ta, která má nejvyšší hodnotu kritéria s nejvyšší váhou.

Pro metody $w_x \in W_A$ platí, že metoda bude zařazena do množiny W_{1A}, pokud splňuje všechna zvolená kritéria (hodnota všech kritérií je nenulová). Metoda je posuzována vzhledem ke konkrétnímu typu prvku. Proto dále musí platit, že metoda dokáže zjistit alespoň některé parametry příslušející k tomuto typu prvku.

Pro každý typ prvku budou evidovány údaje o tom, jaké parametry u tohoto prvku sledujeme a jak vybrané metody dokážou tyto parametry zpracovat. Údaje budou zaznamenány v pomocné matici a hodnoty prvků matice budou tvořeny váženými kritérii. Pro získání hodnoty parametru $a \in A_{smr}$ bude vybrána ta metoda, která pro tento parametr dosahuje nejvyšší hodnoty kritéria. V případě více rovnocenných možností bude zvolena metoda, která poskytuje celkově lepší výsledky. Pro každý typ prvku $t_r \in T$ pak bude existovat množina metod $W_r \subset W_{1A}$, která bude obsahovat metody vhodné pro daný typ prvku.

Metody příslušející do množin W_{1T} a W_{1A} lze použít k analýze dokumentu D. Po jejich aplikaci je výsledkem

- rozdělení dokumentu na stránky P_i, kde $P = \{p_1, \ldots, p_m\}$,
- určení typů jednotlivých prvků ($e\colon P \rightarrow T$),
- zjištění hodnot jejich atributů, tj. naplnění množiny A_{smr} pro všechny prvky.

Výsledkem je *model dokumentu*, který slouží pro další analýzu zaměřenou na hodnocení kvality dokumentu.

ZÁVĚR

Práce se zabývá hodnocením metod pro analýzu a dekompozici dokumentů. Analýza dokumentů je běžně prováděna, sleduje však obecnější cíle. Navržená metodika je oproti tomu orientována do oblasti hodnocení formální kvality dokumentu, což je oblast, která je v literatuře diskutována jen okrajově.

K hodnocení se využívá model dokumentu, který se soustředí pouze na podstatné charakteristiky. Model byl proto navržen tak, aby byl co nejlépe využi-

telný pro hodnocení kvality (především z typografického hlediska) s možností automatizace tohoto procesu. Pravidla pro sazbu dokumentů zdůrazňují principy jednotnosti, čitelnosti, vhodného uspořádání, rovnoměrnosti a vyváženosti obsahu stránky. Proto je nezbytné identifikovat parametry stránky jako celku a také identifikovat prvky stránky.

Rozpoznání jednotlivých prvků na stránce je proveditelné s využitím existujících propracovaných metod pro segmentaci stránky. Klíčový je ovšem typ prvku, tj. zařazení prvku do jedné z vybraných kategorií (tříd). Třídy vycházejí z kategorizace běžně užívané v pravidlech úpravy dokumentů. Zařazení prvku do třídy by mělo být jednoznačné a spolehlivé, protože tyto informace jsou podkladem pro další analýzu a hodnocení.

Pro klasifikaci prvků do jednotlivých tříd lze úspěšně využít neuronovou síť, která splňuje všechny požadavky kladené na rozpoznávání. Dokáže pracovat jako klasifikátor, vyrovná se s nejednoznačností, rozhoduje se na základě vzorů a svá rozhodnutí učením zpřesňuje. Jako nejvhodnější typ sítě se jeví vícevrstvá neuronová síť, bližší určení bude předmětem dalšího výzkumu.

Velmi podstatné jsou rovněž parametry jednotlivých prvků. Každému parametru je přiřazena hodnota, která vstupuje přímo do hodnocení. Je tedy nezbytné vhodně vybrat parametry. Z mnoha možností byla zvolena varianta, kdy jsou parametry závislé na typu prvku a každému typu tedy příslušejí jiné (zato však výstižné) parametry, které charakterizují prvek daného typu. Jednotlivé množiny parametrů (pro každý typ prvku) si nekladou nároky na úplnost, parametry lze dále přidávat a precizovat. Je však třeba najít rovnováhu mezi přesností a náročností hodnocení.

Při získávání parametrů hraje roli výběr vhodné metody. K dispozici je opět řada metod, přičemž nejlépe se uplatňují metody pro analýzu obrazu. Obraz dokumentu, je-li v dobré kvalitě, může podat informace o základních parametrech stránky a jejích prvků (rozměry, umístění…) a lze z něj mj. zjistit intenzitu „zabarvení“ prvku, což je pro hodnocení kvality důležité. Nevýhodou jsou horší předpoklady pro získání charakteristik písma (což je oblast, kde se lépe uplatňují metody vycházející ze strukturovaných formátů dat).

Pro analýzu se v praxi používá řada osvědčených metod, které je třeba ohodnotit z hlediska jejich přínosu pro hodnocení kvality dokumentu. Zkoumají se jak metody pro klasifikaci, tak metody pro zjišťování parametrů, přičemž postup je v obou případech prakticky totožný. Pro ohodnocení přínosu metody se používá skupina kritérií. Kritéria jsou navržena tak, aby zohledňovala využitelnost metody pro hodnocení i efektivní práci s metodou. Množina kritérií si opět neklade nároky na úplnost, kritéria lze dále zpřesňovat a doplňovat, pokud nový stav přispěje ke zkvalitnění analýzy.

Metody, které se účastní hodnocení přínosu metod, jsou ohodnoceny podle vybrané množiny kritérií. Každé kritérium se může skládat z dílčích kritérií – toto zacházení do detailů je důležité pro určení požadavků na metody. Dílčí kritéria jsou popsána hodnotami 0 nebo 1 a tyto hodnoty se vhodně zvolenou metodou kombinují.

Výsledné kritérium je popsáno reálným číslem v rozsahu <0;1>, což vyjadřuje míru jeho splnění. Přidělení ohodnocení však nelze provést automatizovaně, neboť vyžaduje znalost metody. Ohodnocení tedy jednou provede expert a poté je tato metoda připravena pro posuzování z hlediska přínosu pro analýzu.

Hodnocení lze dále zpřesnit a konfigurovat pomocí vektoru vah. Váhy se nastavují jednotlivým kritériím, což umožňuje určit důležitost kritéria, popř. zohlednit tím typ dokumentu. Určením vah se položí důraz na vybrané prvky dokumentu a na jejich parametry. Díky tomu lze nabídnout velké množství alternativ pro hodnocení.

Po splnění popsaných kroků vznikne model reálného dokumentu. Bude obsahovat informace o tom, jaké prvky se na stránce nacházejí, jakého jsou typu a jaké mají parametry; jaké parametry má stránka a jaký je vztah prvků ke stránce. Obdobně lze sledovat vztah jednotlivých stránek a dokumentu. Tyto informace představují podstatný souhrn a jsou dostačující pro hodnocení kvality dokumentu.

Dalším nezbytným zdrojem pro hodnocení je formalizace pravidel určujících kvalitu dokumentu, tj. především typografických pravidel. Data získaná popsaným způsobem i pravidla budou vstupními údaji pro automatizované hodnocení kvality dokumentu. Na tuto práci bude navazovat návrh a hodnocení metodiky pro vlastní hodnocení dokumentu.

SOUHRN

Článek je věnován automatizovanému hodnocení dokumentů z hlediska jejich formální kvality. Formální aspekty přispívají k lepší srozumitelnosti textu, a tedy snazšímu předání informací, proto je nezbytné jejich kontrole a hodnocení věnovat pozornost. Kvalita dokumentu je dána mírou shody reprezentace prvků dokumentu s požadovanými pravidly pro jejich vyjadřování.

Pro hodnocení kvality je navržen model, který vychází z různých úrovní dokumentu a formálně jej popisuje. Dokument je rozdělen na stránky a základem jsou prvky, ze kterých je stránka sestavena. Každému prvku je přiřazen typ, tj. funkce, kterou prvek plní v rámci stránky. Každý prvek je dále popsán množinou parametrů (atributů), které vystihují vlastnosti prvku. Jsou navrženy parametry, které jsou společné pro všechny typy prvků, další parametry jsou závislé na typu prvku.

Parametry prvků jsou vstupem pro hodnocení dokumentu podle typografických pravidel. Je proto zcela nezbytné, aby model věrohodně zobrazoval vlastnosti dokumentu. Na kvalitu modelu mají významný vliv metody, které se používají pro získání informací o prvcích, jejich typech a parametrech. Vzhledem k velkému množství existujících metod je navržena množina kritérií, která popisují požadavky na vhodné metody. Dále je stanoven způsob hodnocení metod podle těchto kritérií včetně možnosti využití vah. Metody, které vyhovují kritériím, lze aplikovat na dokument. Výsledkem je model reálného dokumentu, který je připraven pro automatizované hodnocení na základě typografických pravidel.

kvalita dokumentů, formální kvalita, automatizované hodnocení, model dokumentu, fyzická analýza, logická analýza, typografie

SUMMARY

This paper deals with the automated documents evaluation in terms of their formal quality. Formal aspects make the text clearer and thus facilitate the transmission of information. Therefore, it is necessary to pay attention to their control and evaluation. The quality of the document is given by the degree of agreement between the representations of document objects and the rules required for their expression.

A model is designed for quality evaluation. It results from different document levels and formally describes the document. The document is divided into pages and the basis are the objects from which the page is compiled. A type is assigned to each object, i. e. function which the object fulfils within the page. Each object is described by a set of parameters which reflect the characteristics of the object. Parameters are designed, which are common to all types of elements, other parameters depend on the type of the object.

The parameters of the object are an input for the document evaluation according to the typographical rules. It is therefore absolutely necessary for the model to display the document properties reliably. The quality of the model is significantly influenced by the methods that are used to obtain information about the objects, their types and parameters. Due to the large number of existing methods, a set of criteria is designed, which describe requirements on appropriate methods. The way of evaluation of the methods according to these criteria is also specified, including the use of weights. Methods that meet the criteria can be applied to the document. The result is a model of a real document, which is ready for the automatic evaluation based on the typographical rules.

Článek vznikl v rámci výzkumného záměru MSM 6215648904/03/03/06.

LITERATURA

BAPST, F., INGOLD, R., 1998: Using Typography in Document Image Analysis [online]. [cit. 2006-10-17]. Dostupné z http://citeseer.ist.psu.edu/bapst98using.html.

BEUSEKOM, J. VAN, 2006: Document Layout Analysis [online]. [cit. 2007-02-12]. Dostupné z http://www.iupr.org/~keysers/files/vanBeusekom--DA--Document-Layout-Analysis.pdf.

CATTONI, R. a kol., 1998: Geometric Layout Analysis Techniques for Document Image Understanding: a Review [online]. [cit. 2007-03-27]. Dostupné z http://citeseer.ist.psu.edu/330609.html.

CHAO, H., FAN, J., 2004: Layout and Content Extraction for PDF Documents [online]. [cit. 2007-02-05]. Dostupné z http://www.springerlink.com/content/b928plaetk53ax91/fulltext.pdf.

DOERMANN, D., RIVLIN, E., ROSENFELD, A., 1998: The function of documents [online]. [cit. 2008-01-03]. Dostupné z http://www.cs.technion.ac.il/~ehudr/publications/pdf/DoermannRR98a.pdf

FELICI, J., 2003: The Complete Manual of Typography. Berkeley, Adobe Press. 384 s. ISBN: 0-321-12730-7.

FUSS, C. a kol., 2004: Inferring Structure Information from Typography [online]. [cit. 2007-01-29]. Dostupné z http://www.springerlink.com/content/py61l7j8fufl4t0g/fulltext.pdf.

GORI, M., MARINAI, S., SODA, G., 2003: Artificial Neural Networks for Document Analysis and Recognition [online]. [cit. 2006-10-16]. Dostupné z www.dsi.unifi.it/~simone/ANNxDAR/TR-DSI-01-03.pdf.

HALUZA, P. a kol., 2008: Přístup studentů ke zpracování závěrečné práce. In: Motyčka, A. Informatika XXI/2008. Brno: Konvoj, s. 31–32. ISBN 978-80-7302-151-1.

HARRINGTON, S. a kol., 2004: Aesthetics measures for automated document layout [online]. [cit. 2008-01-28]. Dostupné z http://www.xerox.com/innovation/Aesthetic_Measures.pdf.

HITZ, O., ROBADEY, L., INGOLD, R., 1999: Using XML in Document Recognition [online]. [cit. 2007-02-05]. Dostupné z http://www.science.uva.nl/events/dlia99/final papers/hitz.pdf.

LOVEGROVE, W., BRAILSFORD, D., 1995: Document analysis of PDF files: methods, results and implications [online]. [cit. 2007-01-19]. Dostupné z http://cajun.cs.nott.ac.uk/compsci/epo/papers/volume8/issue2/2point26.pdf.

MA, H., DOERMANN, D., 2005: Font Identification Using the Grating Cell Texture Operator [online]. [cit. 2007-01-05]. Dostupné z http://lampsrv01.umiacs.umd.edu/pubs/Papers/hma-05/hma-05.pdf.

MAO, S., ROSENFELD, A., KANUNGO, T., 2003: Document Structure Analysis Algorithms: A Literature Survey [online]. [cit. 2007-01-19]. Dostupné z http://archive.nlm.nih.gov/pubs/mao/mao03.pdf.

MAŘÍK, V., ŠTĚPÁNKOVÁ, O., LAŽANSKÝ, J. a kol., 1993: Umělá inteligence (1). 1. vyd. Praha: Academia. 264 s. ISBN 80-200-0496-3.

RICHY, H., ANDRÉ, J., 1996: Typographic sheets and structured documents [on-line]. [cit. 2007-01-23] http://cajun.cs.nott.ac.uk/compsci/epo/papers/volume8/issue2/2point5.pdf

RIGAMONTI, M. a kol., 2005: Towards a Canonical and Structured Representation of PDF Documents through Reverse Engineering [online]. [cit. 2007-02-05]. Dostupné z http://diuf.unifr.ch/people/lalanned/Articles/ICDAR05Rigamonti.pdf.

SENNHAUSER, R., 1993: Improving the recognition accuracy of text recognition systems using typographical constraints [online]. [cit. 2007-01-23]. Dostupné z http://cajun.cs.nott.ac.uk/compsci/epo/papers/volume6/issue3/seenhaus.pdf.

SLOCOMBE, D., AMBEKAR, J., 1998: Document Structure Identication: a New Paradigm [online]. [cit. 2007-03-26]. Dostupné z http://www.infoloom.com/gcaconfs/WEB/paris98/slocombe.HTM.

Systémy managementu kvality – Základní principy a slovník. ČSN EN ISO 9000:2006.

TALANDOVÁ, P., 2007: Možnosti kontroly typografické kvality dokumentů. In: Firma a konkurenční prostředí 2007 – Sekce 6: IS/IT a konkurenceschopnost podniků. Brno: MSD, s. r. o., 2007, s. 77–82. ISBN 978-80-86633-88-6.

Adresa

Ing. Petra Talandová, doc. Ing. Jiří Rybička, Dr., Ústav informatiky, Mendelova zemědělská a lesnická univerzita v Brně, Zemědělská 1, 613 00 Brno, Česká republika, e-mail: petra.talandova@pef.mendelu.cz, rybicka@node.mendelu.cz