MANAŽERSKÉ ROZHODOVÁNÍ ZA VYUŽITÍ METOD PRO ZPRACOVÁNÍ DOKUMENTŮ

Hana Kopáčková, Renáta Máchová
Ústav systémového inženýrství a informatiky, Fakulta ekonomicko-správní, UPA

Abstrakt: *Tento příspěvek se zabývá využitím metod pro zpracování dokumentů v rozhodování manažerů. Důraz je kladen na využití nestrukturovaných informací, které v práci manažerů převládají. V příspěvku jsou popsány jednotlivé metody včetně jejich aplikace v praxi.*

Abstract: *This article concerns on usage of text mining methods in managerial decision making. Special concern is given to unstructured information, that is prevalent in managerial practise. Each method is described with its application.*

Klíčová slova: *Rozhodování, textová klasifikace, nestrukturované informace*

Key words: *Decision making, text classification, unstructured information*

1. Úvod

Současná doba je přesycená nejrůznějšími informacemi, mezi nimiž je velice složité se pohybovat. Umět zacházet se získanými daty efektivně a v rámci prosperity firmy je nezbytně důležité. Dnes jsou informace dostupné různými způsoby z velkého množství různých zdrojů. Informační zdroje jsou rozptýleny a každý den vznikají nové. Dnešní dostupné informační zdroje zahrnují informace z institucí veřejné správy, digitálních knihoven, nezávislých zpravodajských agentur, nevládních organizací či nezávislých odborníků poskytujících různé služby. Právě velké množství informačních zdrojů, různé úrovně jejich přístupnosti, důvěryhodnosti a spojených nákladů způsobují značný problém při plánování pořizování informací. Proces tvorby informace, jejího získání, uložení (např. do databáze informačního systému), její analýzy a distribuce uživatelům vyžaduje určitý výpočetní výkon a čas. Informace však může být v rozhodovacím procesu využita jen pokud je nalezena, zpracována a dále distribuována v krátkém čase. Manažeři se musí rozhodovat velmi často; některá rozhodnutí jsou velmi významná, takže rychlý přístup ke všem potřebným informacím je pro ně naprostou nezbytností. [1, 2]

Jednou z nejvýznamnějších aktivit manažerů je umět se správně rozhodnout. Kvalita a výsledky manažerského rozhodnutí totiž ovlivňují fungování a prosperitu jednotlivých společností a firem a právě nekvalitní rozhodování může být příčinou jejich neúspěchů. [3]

Kvalita manažerského rozhodování je ovlivňována řadou činitelů, k nimž lze zařadit zejména vědomosti, schopnosti a dovednosti manažerů, druh řešeného rozhodovacího problému, vybavenost moderními technickými prostředky řízení a komunikace, časovým horizontem rozhodování a jeho požadovanými návaznostmi, změnami ve vnějším a vnitřním prostředí managementu a podnikatelským rizikem, které provází v tržní ekonomice všechny podnikatelské aktivity. Z uvedených informací je patrné, že manažerské rozhodování je velmi komplikované a zahrnuje v sobě kombinaci velmi různorodých a dynamicky se měnících procesů jež mají vliv na výsledné rozhodnutí. [4]

Existuje řada přístupů k rozhodování, které závisí především na charakteru problémů, čase a na schopnostech manažera. Rozhodování lze považovat spíše za prostředek než požadovaný výsledek. Je to proces, s jehož pomocí chce manažer dosáhnout požadovaného stavu. Výsledkem rozhodování je rozhodnutí. Každé rozhodnutí je výsledkem dynamického procesu, který je ovlivňován mnoha faktory (organizační prostředí, manažerské dovednosti, motivace,...). Čím více se jedná o ojedinělý problém a čím více budou výsledky ovlivňovány

neurčitostí, tím bude realizován rozhodovací proces kompletněji. Z toho je jasné, že rozhodovací proces se využívá především pro neprogramovaná rozhodování (problém je nový, tzn. specifický a do určité míry neopakovatelný, který není nijak vázán s problémy minulými).[5, 6, 7]

2. Strukturované versus nestrukturované informace

Schopnost získat z dostupných dat více informací, než dokáží jiní, je nepochybně v přímé úměře ke kvalitě rozhodnutí. Takto se vytváří silný prvek konkurenční výhody. Cest, jak získat z údajů to důležité, je mnoho, ale v popředí vždy stojí metody statistické analýzy dat. Statistické zpracování dat vyžaduje strukturovanost údajů a jejich měřitelnost. V práci manažera je možno najít mnoho informací v této strukturované podobě např. tabulky, grafy, databáze, výstupy z dotazníků s uzavřenými odpověďmi, atd. Výhodou tohoto přístupu je jednoduchost zpracování, neboť statistické metody, které pracují se strukturovanými daty jsou dobře známé a bývají součástí aplikačního software. Mezi nevýhody patří zejména omezení vyjadřovacího potenciálu zdroje informací. Další nevýhodou je nutnost dopředu definovat strukturu požadované informace, což v některých případech není vůbec možné (např. při sledování aktualit z oblasti finančních trhů). Tyto nedostatky je možno odstranit použitím nestrukturovaného dokumentu ve formě souvislého textu.

V mnoha oborech se můžeme setkat s textovými dokumenty, které představují jistou formu předpisu nebo doporučení. Mohou to být např. lékařská doporučení pro léčbu určité choroby, stavební standardy a normy či právní spisy. Z pohledu obsažených znalostí se tyto dokumenty vyznačují určitými specifickými rysy. Především zahrnují ucelený soubor doménových znalostí o dané problematice a dále na sebe vážou jistou všeobecnou platnost. Informace obsažené v této podobě mohou být zpracovány formou klasifikace a zařazení do tříd podle obsahové shody. Tento postup je nazýván jako textová klasifikace. Dostupné metody pro textovou klasifikaci se odlišují dle použité metody učení. Při použití metod učení s učitelem je nejprve definován počet a označení tříd. Základem je vytvoření trénovacích množin, které obsahují dokumenty správně přiřazené a díky těmto příkladům je pak možno zařazovat dokumenty nové. [8, 9, 10] Metoda učení bez učitele se vyznačuje tím, že nejsou k dispozici žádná požadovaná výstupní data, učení tedy probíhá pouze na základě vstupních dat. Účelem těchto metod je rozpoznat v datech jisté pravidelnosti, aniž by zde byla zpětná informace o tom, kde a jak tyto pravidelnosti hledat. Z tohoto důvodu je učení bez učitele používáno pro řešení úloh jako detekce podobnosti nebo shluková analýza. [11]

Popisované metody jsou velmi úspěšně aplikovány při zařazování elektronických dokumentů do témat v rámci diskusních skupin, při filtrování nežádoucích stránek a zpráv nebo při vytváření katalogizovaného vyhledávače. V oblasti manažerského rozhodování se však tyto metody nepoužívají, ačkoliv by jejich použití mělo velký vliv na rychlost a kvalitu rozhodovacího procesu na všech úrovních managementu. V současné době neexistují srovnání jednotlivých metod strojového učení při zpracování strukturovaných i nestrukturovaných textových dokumentů a zároveň hodnocení přínosu pro manažerské rozhodování. Základní otázkou prozatím zůstává: „Je možno použitím metod pro zpracování dokumentů zjednodušit proces rozhodování?“

3. Metody pro zpracování dokumentů

Obor, který se zabývá zpracováním dokumentů a nalezením informací v nich obsažených se momentálně zahrnuje pod pojem text mining. Text mining zahrnuje nejen dolování v dokumentu, ale také dolování v množině dokumentů. Níže jsou popsány nejdůležitější metody.

Kategorizace (Categorization) – představuje zařazení testovacích dokumentů s podobným tématem do předdefinovaných tříd na základě zařazení trénovacích dokumentů.

Shlukování (Clustering) – sdružuje podobné dokumenty do shluků bez předchozího trénování.

Sumarizace (Summarization) – zestručňuje text na několik vybraných vět.

Extrakce (Extraction) – vybírá z dokumentu pouze požadované části textu (názvy firem, ceny...)

Selekce informací (Information retrieval)– vyhledává daný dokument na základě klíčových slov.

Vizualizace (Visualisation) – graficky prezentuje hledaná data, takže je jednodušší porozumění vztahům mezi nimi.

Jednotlivé metody představují nástroje pro práci s textovou informací a pokud jsou vhodně použity mohou usnadnit manažerům práci. Hledání informací a jejich zpracování je totiž časově velmi náročné.

3.1. Textová kategorizace

Problematika kategorizace je zaměřena na klasifikaci dokumentů do předem definovaných tříd, ovšem občas je tento problém složitý a v některých případech dokonce nemá řešení vůbec. Základem je vytvoření trénovacích množin, které obsahují dokumenty správně přiřazené a díky těmto příkladům je pak možno zařazovat dokumenty nové. Tato úloha, která se nazývá učení s učitelem (supervised learning) [12] je použita při úlohách rozpoznávání a strojovém učení.

3.2. Shlukování

Je souhrnný název pro celou řadu výpočetních postupů, jejichž cílem je rozklad daného souboru na několik relativně homogenních podsouborů (shluků) a to tak, aby jednotky (objekty) uvnitř jednotlivých shluků si byly co nejvíce podobné a jednotky (objekty) patřící do různých shluků si byly podobné co nejméně. Při tom každá jednotka je popsána skupinou znaků (proměnných). Výsledky analýzy závisí na volbě proměnných, zvolené míře vzdálenosti mezi objekty a shluky a na zvoleném algoritmu výpočtu. Na závěr shlukovací analýzy se proto provádí charakterizace (popis) jednotlivých tříd (tj. shluků) a interpretace. Shlukovací metody jsou úspěšné především v situacích, kdy objekty mají tendenci se seskupovat do přirozených tříd, než v případě náhodného rozmístění objektů v atributovém prostoru.

3.3. Sumarizace

Automatická sumarizace dokumentu umožňuje uživateli v krátkém čase porozumět obsahu daného dokumentu. Systém automatické sumarizace ze vstupního textu vyrobí souhrn jeho důležitých částí. Rozeznáváme dva hlavní typy sumarizace:
Sumarizace extrakcí – Souhrn je vyjmut z původního textu pomocí statistických principů nebo za pomocí heuristických metod nebo pomoci obou. Vyjmuté části již nejsou syntakticky ani obsahově měněny.

Obsahová sumarizace – Souhrn je interpretací původního textu. Vzniká přepsáním původního textu tak, že nahrazuje příliš dlouhé části textu jejich kratšími interpretacemi. Například větu „Pěstovala na poli papriku, rajčata a okurky.“ může být nahrazena větou „Pěstovala na poli zeleninu.“

3.4. Extrakce

Indukce pravidel, která extrahují položky daného typu z určeného dokumentu, speciálně se tato metoda používá pro zpracování HTML a XML stránek. Tedy takových, které používají značkovacího jazyka. Samotný proces extrakce vyžaduje trénovací stránky, kde je ukázáno jaké údaje extrahovat (např. vše co je mezi značkami <a> ....). Ve chvíli, kdy je natrénovaný automat uložen, může denně extrahovat údaje z dané stránky. Bohužel obvykle

přestává fungovat při změně designu stránky, a proto je vhodné používat nástroj pro kontrolu HTML kódu.

3.5. Selekce informací

Selekce informací zahrnuje dolování v množině dokumentů, tedy hledání konkrétního dokumentu podle klíčových slov. Tuto metodu používají vyhledávače např. Seznam,Google, atd. Uživatel zadá klíčová slova do formuláře a jako výsledek obdrží seznam relevantních dokumentů. Problémem ovšem bývá, že uživatel často není spokojen s nalezenými dokumenty. Při posuzování kvality libovolného systému pro vyhledávání informací hraje hlavní roli relevance, která se definuje jako vlastnost vztahu mezi dotazem uživatele a jednotlivým dokumentem jako prvkem množiny všech nalezených dokumentů. Na základě posouzení relevance výsledků se pak odvozují dvě, velmi často používané míry hodnocení efektivnosti vyhledávání: přesnost (precision) a úplnost (recall) [13].

3.6. Vizualizace

Tato metoda je použitelná pro všechny předchozí jako doplňková. Například při shlukování ve 2-D prostoru ukazuje, jaké je rozložení dokumentů a jakým způsobem byly shluky vytvořeny. Můžeme proto lépe posoudit, zda bylo provedeno správně. Stejně tak při kategorizaci metodou rozhodovacích stromů bývá používána vizualizace.

3.7. Aplikace jednotlivých metod při rozhodování

V rozhodovacím procesu lze použít buď jednotlivé metody, nebo jejich kombinaci. Například můžeme kombinovat selekci informací s textovou kategorizací. Tato kombinace umožňuje nalézt opravdu relevantní dokumenty v co nejkratším čase. Ukažme si aplikaci této kombinace na příkladu. Firma XY chce rozšířit portfolio vyráběných produktů o technické oblečení z materiálů specifických vlastností. Chce se tedy dozvědět maximum informací o použitých materiálech, výrobcích, oblíbenost a zkušenosti s jednotlivými materiály, atd. Vytvoří si tedy klasifikátor s trénovací množinou dokumentů, které nejlépe vystihují dané téma jako přípravu pro textovou klasifikaci. Pokud bychom chtěli v tomto bodě prověřit všechny dokumenty z Internetu zda odpovídají požadavkům klasifikátoru, trvala by celá akce nejspíše roky. Proto je vhodné nejprve použít metodu selekce informací ve spolupráci s vhodným vyhledávačem. Zde zadat několik klíčových slov a získat množinu dokumentů, které by mohly být relevantní. Nicméně zkušenosti ukazují, že 70 % vyhledaných dokumentů nesplňuje požadavky na ně kladené. Jejich prohlížením bychom ztráceli mnoho času, a proto je vhodné v této fázi využít textové kategorizace, která lépe zaručí tématickou shodu. Celkově lze tedy shrnout, že kombinace uvedených metod snižuje časové nároky a zvyšuje přesnost.

Jinou kombinací může být použití textové kategorizace a extrakce informací. První metoda určí zda se jedná o dokument relevantní a druhá metoda vybere určitou relevantní část textu. Použití můžeme najít například při porovnání cenových nabídek, které přišly e-mailem. Nejprve je třeba vybrat dokument, který obsahuje cenovou nabídku a v druhé fázi je z dokumentu vyfiltrována informace o jakou firmu se jedná a jakou cenu nabízí. Tyto informace mohou být ukládány například do přehledné tabulky.

Na závěr je možno říci, že metody pro zpracování dokumentů opravdu mohou ušetřit čas a tím zjednodušit proces rozhodování.

4. Poděkování

Tento příspěvek vznikl za finanční podpory ze strany Grantové agentury České republiky, grantové číslo GACR 402/05/P155

5. Literatura

1. Oates, T., Prasad, N., Lesser, V., Cooperative information gathering: A distributed problem solving approach. Technical Report UMASS 94-66, Dept. of Computer Science, Univ. of Massachusetts, 1994
2. Zilberstein, S., Lesser, V., Intelligent information gathering using decision models. CS Technical Report 96–35, Univ. of Massachusetts, 1996
3. Hindls, R., Hronová, S., Novák, I.: Analýza dat v manažerském rozhodování. Grada, Praha 1999
4. Svoboda, E. Moderní přístupy ve strategickém rozhodování managementu v podnicích s agrárním předmětem činnosti. [online] URL: <http://apxi.pef.czu.cz/pdf/ap-180.pdf> [cit. 2004-10-2]
5. Crainer, S.: Moderní management. Základní myšlenkové směry, Management Press, Praha 2000
6. Bowman C.: Strategický management, Grada, Praha 1996
7. Stead W.E., Stead J.G.: Management pro malou planetu, Strategická rozhodování a životní prostředí, G plus G, 1998
8. Quinlan, J. R. Induction of Decision Trees. *Machine Learning*, 1986, vol. 1, no.1.
9. Quinlan, J. R. *C4.5:Programs for Machina Learning*. Morgan Kaufman, 1993.
10. Mitchell, T. M. *Machine Learning*. New York: McGraw Hill, 1996.
11. Agrawal R., Imielinski T., Swami A.:Mining association rules between sets of items in large databases.Proc. of ACM SIGMOD Conference on Management of Data, 1993.
12. Quinlan J. R.: C4.5:Programs for Machina Learning, Morgan Kaufman, 1993.
13. Y. Yang. An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, vol. 1, num. 2, 1999.

Kontaktní adresy:
Ing. Hana Kopáčková, Ph.D.
Ústav systémového inženýrství a informatiky, Fakulta ekonomicko-správní,
Univerzita Pardubice, Studentská 84, 532 10 Pardubice,
tel.: 466 036 245
e-mail: hana.kopackova@upce.cz

Ing. Renáta Máchová, Ph.D.
Ústav systémového inženýrství a informatiky, Fakulta ekonomicko-správní,
Univerzita Pardubice, Studentská 84, 532 10 Pardubice,
tel.: 466 036 074
e-mail: renata.machova@upce.cz