POSUDEK OPONENTA DIPLOMOVÉ PRÁCE

Jméno diplomanta: Bc. Viktor Nekvapil

Název práce: Data Mining in Customer Relationship Management: The Case of Major Logistic Company

Hledisko 1: náročnost práce

Náročnost tématu na teoretické znalosti	☐ vysoká	☒ průměrná	☐ nižší
Náročnost tématu na praktické zkušenosti	☒ vysoká	☐ průměrná	☐ nižší
Náročnost tématu na na rozsah a úroveň podkladových materiálů	☒ vysoká	☐ průměrná	☐ nižší

Hledisko 2: splnění věcných požadavků

Stupeň splnění cíle práce	☐ vynikající	☒ solidní	☐ vyhovující	☐ nevyhovující
Hloubka provedené analýzy ve vztahu k tématu	☒ vynikající	☐ solidní	☐ vyhovující	☐ nevyhovující
Aktuálnost uvedených údajů	☒ vynikající	☐ solidní	☐ vyhovující	☐ nevyhovující
Adekvátnost použitých metod (pramenů)	☐ vynikající	☒ solidní	☐ vyhovující	☐ nevyhovující
Vlastní přínos k řešené problematice	☒ vynikající	☐ solidní	☐ vyhovující	☐ nevyhovující
Stupeň realizace navrhovaných řešení (programy ap.)	☐ vynikající	☐ solidní	☐ vyhovující	☐ nevyhovující
Využitelnost výsledků práce v praxi (teorii)	☒ vysoká	☐ průměrná	☐ nižší	

Hledisko 3: formální náležitosti

Logická stavba práce	☐ vynikající	☒ solidní	☐ vyhovující	☐ nevyhovující
Formální bezchybnost textu	☒ vynikající	☐ solidní	☐ vyhovující	☐ nevyhovující
Grafická úprava práce	☒ vynikající	☐ solidní	☐ vyhovující	☐ nevyhovující
Adekvátnost a srozumitelnost závěrů	☒ vynikající	☐ solidní	☐ vyhovující	☐ nevyhovující
Práce s literaturou (citace)	☐ vynikající	☒ solidní	☐ vyhovující	☐ nevyhovující

Připomínky a otázky vyžadující doplnění:[1]
viz komentář na dalších stránkách.

[1] Podle potřeby rozepište na dalších stranách.

Práci doporučuji k obhajobě.

Navržená výsledná známka: výborně

V Praze dne 20. 7. 2012

Jméno: doc. Ing. Milan Šimůnek, Ph.D. Podpis:
……

Diplomová práce: *Data Mining in Customer Relationship Management: The Case of a Major Logistic Company*
Autor: *Bc. Viktor Nekvapil*

Diplomová práce popisuje v anglickém jazyce data-miningovou analýzu obchodních případů významné zásilkové firmy.

Text o rozsahu 76 stran je rozdělen do tří nestejně dlouhých sekcí, seznamu použité literatury a pěti příloh. Po úvodu následuje první sekce s krátkým, jednostránkovým popisem systému LISp-Miner, který byl následně použit při analýze. Následuje druhá sekce s podrobným popisem provedených analýz strukturovaným podle metodiky CRISP-DM. Byly provedeny celkem dvě iterace, z nichž druhá vzala v úvahu reakce doménových expertů na výsledky dosažené v první iteraci. Následuje klíčová, třetí sekce DP, ve které autor popisuje získané zkušenosti a doporučení pro případné další analýzy takového druhu. Kromě shrnutí výsledků práce následuje ještě pět příloh poukazujících na kroky, které bylo nutné v rámci analýzy nutné vykonat.

Na práci jsou vidět jak teoretické, tak i praktické zkušenosti autora, množství odvedené práce i těžkosti, se kterými se v praxi musel potýkat. Důležité jsou vlastní přínosy autora ve formě uvedených zkušeností a doporučení. I přes níže uvedené připomínky práci jednoznačně doporučuji k obhajobě.

Vlastní přínos a práce s literaturou

Vlastní přínos autor je zejména ve třetí sekci, kde jsou velmi cenné zkušenosti s prováděním data-miningové analýzy v prostředí komerční firmy a i doporučení pro možné rozšíření modulu *LM DataSource* ve fázi kategorizace dat. I samotný popis obou iterací analýzy je však pěkně zpracovaný a dává další podněty k zamyšlení a možným směrům jejího dalšího rozšiřování.

Práce s literaturou je adekvátní. V případě, že autor přebírá informace nebo se odvolává, tak je citace uvedena. Na konci práce je uveden seznam použité literatury. Většinu práce však tvoří autorem prováděná analýza, kde nebylo možné/nutné se na literaturu odvolávat.

Věcné připomínky

V diagramu na obrázku 3 (s. 11) postrádám šipky vedoucí z datový úložišť (Suspect, Development lead, Opportunity, Activity). Bylo by velmi zvláštní, kdyby se do nich data pouze ukládala a neexistoval proces, který by z nich naopak četl.

V diagramu na obrázku 4 (s. 13) nebylo nutné proces uzavírání případů (Future oportunity × Closed lost) kreslit vícekrát. Naopak je to matoucí, protože se pak zdá, že jde o dva různé procesy. Duplicitně se procesy zobrazují pouze v případě, že na obrázku není dost místa, resp. by vzniklo příliš mnoho křížení čar. To však není tento případ. Opět se žádná informace nenačítá z Opportunity a i zde je použita nestandardní syntaxe pro data-flow diagramy (více viz níže), u které navíc není vysvětlen symbol ⊗.

Opravdu je maximální počet zásilek pro jednoho zákazníka 1,8 milionu? Jde-li o data za cca 7 let, tak je to cca 260 tis. zásilek za rok, tedy přes 700 denně. Možné to je, ale v takovém případě stálo za komentář v části „Příprava dat“.

Je škoda, že nebyla řešena i druhá analytická otázka, i když chápu, že nutná příprava dat by vyžadovala ještě dost práce a přímočaré není ani využití informací z vytvořených sekvencí – viz doplňující otázka.

Patrně zejména kvůli přílišné zaneprázdněnosti doménových expertů není v práci více rozebrán třetí cíl – návrh jednoduchého a srozumitelného způsobu prezentace výsledků

procedur systému LISp-Miner majitelům dat. Vyřešení tohoto problému by mělo významný dopad a je škoda, že se v tomto ohledu nepodařilo dosáhnout více.

Struktura práce

Struktura práce mi přijde poněkud nevyvážená, protože první sekce s popisem použitého systému svým rozsahem jen velmi málo přesahuje jednu stránku A4. Bylo by například vhodné více rozvést rozšířenou syntaxi jak samotných asociačních pravidel, tak i jejich rozšíření v procedurách *SD4ft-Miner* a *Ac4ft-Miner*. Protože však autor již svou bakalářskou prací dostatečně prokázal, že metodě GUHA rozumí a ovládá i samotný systém LISp-Miner, tak nepovažuji krátkou první sekci za zásadní problém.

Za zbytečné považuji číslování nadpisů až do čtvrté úrovně (viz např. s. 18 – 1.2.1.2).

Srozumitelnost textu

Na práci je místy vidět, že autor studoval původně jiný obor než informatika. Nejvýraznější je to u diagramů (s. 11, s. 13) a relačního schématu (s. 16). V obou případech existují pro popis těchto vztahů zažité typy diagramů a konvence, které se v nich používají – např. v relačním schématu (resp. ERA diagramu) určují šipky a jejich směr kardinalitu vztahů. Podobně jsou v *data-flow* diagramech již navržené symboly pro datová úložiště, procesy atp. Stačilo použít některý z četných CASE nástrojů, které nabízejí kreslení těchto diagramů. Kreslení relačních schémat bývá i součástí DBMS. Určitě není vhodné vymýšlet vlastní názvy („*a kind of mind-map*“) a symboly. Dále, hned v první větě abstraktu např. hovoří o LISp-Mineru jako o „open-source“ systému – to je nepřesné. LISp-Miner je možné nazvat free-ware (nabízí se zdarma), ale zdrojový kód zatím zveřejněn nebyl, a není tedy možné mluvit o „open-source“.

Domnívám se, že vzhledem k rozsahu odvedené práce, je možné v diplomové práci výše uvedené tolerovat. V případném doktorském studiu a před psaním disertační práce by však měl autor teoretické znalosti ještě doplnit.

Za velmi nešťastnou považuji zkrácenou formulaci analytické otázky: „*Which combinations of salesman and lead source have the highest revenue/closing ratio/share of closed won opportunities?*“ (s. 15). Vzhledem k tomu, že jde o jeden z nejdůležitějších textů v celé práci, tak je na škodu jakákoliv nejednoznačnost, kterou vzbuzuje – myslí se vše najednou? Nebo jsou tyto tři možnosti ekvivalentní? Nebo (jak je asi myšleno) jde o **TŘI** různé otázky? Vzhledem k důležitosti otázky by bylo daleko lepší ji napsat jako tři otázky (jednu pro „*the highest revenue*“, druhou pro „*closing ratio*“ a třetí pro „*share of closed won opportunities*“.

Občas se stává, že autor opakuje doslovně již jednou uvedené věty, ale nejde o tak častý jev. Zbytečně dlouhý je také nadpis třetí sekce, který je hned v následující větě znovu celý zopakován (s. 50).

Na začátku práce používá autor pro pojem „data“ jednotné číslo (např. s. 16). Osobně bych dal přednost množnému, tedy „*the data are produced*“ místo „*the data is produced*“, ale zdá se, že jde o novodobý trend a ani přirození mluvčí se nedokáží shodnout. Co je však špatně, že dále v textu je použito střídavě množné i jednotné číslo (např. s. 54 „*the data include*“ a s. 55 „*the data come**s***“).

V metodě GUHA se pro výsledek data-miningové analýzy používá pojem „hypotéza“. Proto mi nepřijde vhodné používat jej v textu i jinde, kde nahrazuje slovo „domněnka/předpoklad“ (s. 18) – není to nic zásadního, ale použití pojmu může čtenáře zmást.

Formální úprava

Formální úprava práce je dobrá a oproti jiným závěrečným pracím obsahuje relativně málo gramatických chyb a překlepů, uvážíme-li navíc, že je napsána v anglickém jazyce. Je však

škoda, že hned v českém abstraktu je zapomenuté „v“ na konci řádku, jeden překlep a dvakrát slovo od slova stejná věta.

Autor je nedůsledný v zakončování vět v číslovaných/bodových seznamech. Ve většině případů není na konci bodu tečka (případně čárka nebo středník), ale někdy tečku uvede. Konečně bych ještě doporučil používání záhlaví stránek, kde by byl uveden název kapitoly nebo přílohy. To by usnadnilo orientaci textu, zejména při hledání odkazovaných příloh.

Další poznámky jsem vyznačil přímo do textu, i když například v případě použití určitých a neurčitých členů jde spíše o návrhy.

Dotazy k obhajobě

- Jak by chtěl autor předzpracovat sekvence aktivit, aby informace z nich bylo možné použít v některé z procedur systému LISp-Miner (aby např. byla rozpoznaná podobnost dvou sekvencí, které mají shodných posledních X aktivit? Nebo lišících se pouze v jedné aktivitě navíc – např. v druhé sekvenci na Y-té pozici? atp.)
- Na straně 52 autor navrhuje heuristický přístup ke kategorizaci. Nešlo by obdobného výsledku dosáhnout tak, že by byly automaticky vygenerovány kratší ekvidistantní intervaly a bylo ponecháno na *4ft-Mineru*, aby jejich spojováním našel vhodné meze, navíc s požadovanou minimální frekvencí?