VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

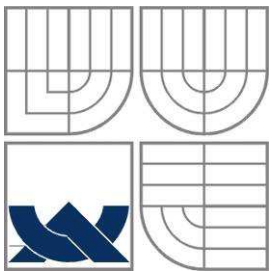
DOLOVÁNÍ Z DAT V PROSTŘEDÍ INFORMAČNÍHO SYSTÉMU K2

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

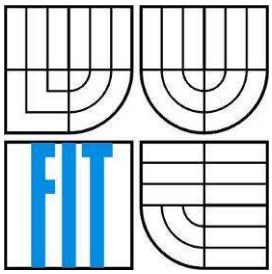
AUTOR PRÁCE
AUTHOR

BC. PETR FIGURA

BRNO 2007



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

DOLOVÁNÍ Z DAT V PROSTŘEDÍ INFORMAČNÍHO SYSTÉMU K2

DATA MINING IN K2 INFORMATION SYSTEM

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — BC. PETR FIGURA
AUTHOR

VEDOUCÍ PRÁCE — DOC. ING. JAROSLAV ZENDULKA, CSC.
SUPERVISOR

BRNO 2007

Zadání diplomové práce/3478/2006/xfigur00

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav informačních systémů Akademický rok 2006/2007

Zadání diplomové práce

Řešitel: **Figura Petr, Bc.**
Obor: Informační systémy
Téma: **Dolování z dat v prostředí informačního systému K2**
Kategorie: Databáze

Pokyny:

1. Seznamte se s problematikou získávání znalostí z databází.
2. Podrobně se seznamte s informačním systémem K2 a strukturou jeho databáze.
3. Navrhněte, které moduly informačního systému a které části databáze jsou vhodné pro získávání znalostí z dat v nich uložených. Dále vytipujte několik analytických úloh pro tato data a určete odpovídající typy dolovacích úloh.
4. Po dohodě s vedoucím zvolte jeden z typů dolovacích úloh (např. asociační pravidla) a nastudujte algoritmy jejich řešení.
5. Navrhněte a implementujte modul systému K2 na podporu vybraného typu dolovací úlohy. Funkčnost ověřte na vhodně zvoleném vzorku dat.
6. Zhodnoťte dosažené výsledky. Diskutujte další možný rozvoj systému.

Literatura:

1. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Second Edition. Elsevier Inc., 2006, 770 p., ISBN 1-55860-901-3.
2. Berka, P.: Dobývání znalostí za databází. Academia, 2003, 366 s., ISBN 80-200-1062-9.

Při obhajobě semestrální části diplomového projektu je požadováno:

- Body 1 až 4.

Podrobné závazné pokyny pro vypracování diplomové práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva diplomové práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap, které byly vyřešeny v rámci ročníkového a semestrálního projektu (30 až 40% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním paměťovém médiu (disketa, CD-ROM), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Zendulka Jaroslav, doc. Ing., CSc.**, UIFS FIT VUT
Konzultant: Šarman Zdeněk, Mgr., K2 atmitec
Datum zadání: 28. února 2006
Datum odevzdání: 22. května 2007

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav informačních systémů
612 66 Brno, Božetěchova 2

doc. Ing. Jaroslav Zendulka, CSc.
vedoucí ústavu

LICENČNÍ SMLOUVA
POSKYTOVANÁ K VÝKONU PRÁVA UŽÍT ŠKOLNÍ DÍLO

uzavřená mezi smluvními stranami

1. Pan

Jméno a příjmení: **Bc. Petr Figura**
Id studenta: 49256
Bytem: Lískovec 32, 738 01 Frýdek-Místek
Narozen: 13. 09. 1982, Frýdek-Místek
(dále jen "autor")

a

2. Vysoké učení technické v Brně

Fakulta informačních technologií
se sídlem Božetěchova 2/1, 612 66 Brno, IČO 00216305
jejímž jménem jedná na základě písemného pověření děkanem fakulty:

.....
(dále jen "nabyvatel")

Článek 1
Specifikace školního díla

1. Předmětem této smlouvy je vysokoškolská kvalifikační práce (VŠKP):
diplomová práce

Název VŠKP: Dolování z dat v prostředí informačního systému K2
Vedoucí/školitel VŠKP: Zendulka Jaroslav, doc. Ing., CSc.
Ústav: Ústav informačních systémů
Datum obhajoby VŠKP: .....

VŠKP odevzdal autor nabyvateli v:

tištěné formě počet exemplářů: 1
elektronické formě počet exemplářů: 2 (1 ve skladu dokumentů, 1 na CD)

2. Autor prohlašuje, že vytvořil samostatnou vlastní tvůrčí činností dílo shora popsané a specifikované. Autor dále prohlašuje, že při zpracovávání díla se sám nedostal do rozporu s autorským zákonem a předpisy souvisejícími a že je dílo dílem původním.
3. Dílo je chráněno jako dílo dle autorského zákona v platném znění.
4. Autor potvrzuje, že listinná a elektronická verze díla je identická.

Článek 2
Udělení licenčního oprávnění

1. Autor touto smlouvou poskytuje nabyvateli oprávnění (licenci) k výkonu práva uvedené dílo nevýdělečně užít, archivovat a zpřístupnit ke studijním, výukovým a výzkumným účelům včetně pořizování výpisů, opisů a rozmnoženin.
2. Licence je poskytována celosvětově, pro celou dobu trvání autorských a majetkových práv k dílu.
3. Autor souhlasí se zveřejněním díla v databázi přístupné v mezinárodní síti:
   - ☐ ihned po uzavření této smlouvy
   - ☐ 1 rok po uzavření této smlouvy
   - ☒ 3 roky po uzavření této smlouvy
   - ☐ 5 let po uzavření této smlouvy
   - ☐ 10 let po uzavření této smlouvy

   (z důvodu utajení v něm obsažených informací)
4. Nevýdělečné zveřejňování díla nabyvatelem v souladu s ustanovením § 47b zákona č. 111/ 1998 Sb., v platném znění, nevyžaduje licenci a nabyvatel je k němu povinen a oprávněn ze zákona.

Článek 3
Závěrečná ustanovení

1. Smlouva je sepsána ve třech vyhotoveních s platností originálu, přičemž po jednom vyhotovení obdrží autor a nabyvatel, další vyhotovení je vloženo do VŠKP.
2. Vztahy mezi smluvními stranami vzniklé a neupravené touto smlouvou se řídí autorským zákonem, občanským zákoníkem, vysokoškolským zákonem, zákonem o archivnictví, v platném znění a popř. dalšími právními předpisy.
3. Licenční smlouva byla uzavřena na základě svobodné a pravé vůle smluvních stran, s plným porozuměním jejímu textu i důsledkům, nikoliv v tísni a za nápadně nevýhodných podmínek.
4. Licenční smlouva nabývá platnosti a účinnosti dnem jejího podpisu oběma smluvními stranami.

V Brně dne: .....

....

Nabyvatel Autor

Abstrakt

Tento projekt vznikl jako firemní zadání společnosti K2 atmitec Brno s.r.o. Jeho výsledkem je modul pro dolování dat v prostředí informačního systému K2. Navrhovaný modul implementuje asociační analýzu nad daty datového skladu informačního systému K2. Analyzovaná data obsahují informace z prodejů evidovaných v informačním systému K2. Modul tedy implementuje analýzu nákupního košíku.

Klíčová slova

dolování dat, získávání znalostí z databází, data, databáze, výběr dat, čištění dat, transformace dat, ETL, dolovací modul, datový sklad, analýza nákupního košíku, asociační analýza, asociační pravidlo, kandidátní množina, shluková analýza, klasifikace a predikce, hvězdicové schéma, K2 skript, manažerské vyhodnocování, informační systém, K2

Abstract

This project was originated by K2 atmitec Brno s.r.o. company. The result is data mining module in K2 information system environment. Engineered data module implements association analysis over the data of K2 information system data warehouse. Analyzed data contains information about sales filed in K2 information system. Module is implementing consumer basket analysis.

Keywords

data mining, knowledge discovery in databases, data, database, data selection, data cleaning, data transformation, ETL, data mining module, data warehouse, market basket analysis, association analysis, association rule, candidate set, cluster analysis, classification and prediction, star schema, K2 script, manager evaluation, information system, K2

Citace

Figura Petr: Dolování z dat v prostředí informačního systému K2. Brno, 2007, diplomová práce, FIT VUT v Brně.

Dolování z dat v prostředí informačního systému K2

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením docenta Jaroslava Zendulky, Ing., CSc.
Další informace mi poskytl Mgr. Zdeněk Šarman.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Jméno Příjmení
Datum

Poděkování

Děkuji docentu Jaroslavovi Zendulkovi, Ing., CSc. za rady a konzultace na projektu. Děkuji Mgr. Zdeňkovi Šarmanovi za rady a doporučení během realizace projektu.

Obsah

1 Úvod

Množství dat uložených v databázích se v současné době velice zvyšuje, ovšem množství užitečných informací (znalostí) z těchto dat se z důvodu velkého objemu dat snižuje. K řešení tohoto problému se v posledních letech rozvinula oblast získávání znalostí z databází. Nejedná se pouze o výzkumnou oblast, ale výsledky poskytnuté pomocí těchto metod se uplatňují i v reálném světě. Pomocí metod získávání znalostí z databází je možné snižovat cenu produktů, optimalizovat obchodní operace, nebo například předpovídat trendy cen na burze. Takovéto informace již není možné jednoduše získat pomocí strukturovaného nebo dotazovacího jazyka [2].

Tento dokument popisuje a vysvětluje proces analýzy, návrhu a tvorby modulu pro získávání znalostí z databází. Algoritmy získávání znalostí pracují s daty uloženými v datovém skladu, přičemž export dat do datového skladu provádějí skripty pro OLAP analýzu informačního systému K2 společnosti K2 atmitec s.r.o. Projekt vznikl jako rozšíření stávajícího informačního systému K2 v brněnské firmě K2 atmitec Brno s.r.o. ve spolupráci s firmou K2 atmitec s.r.o. Navazuje na mou bakalářskou práci Export dat z informačního systému K2 společnosti Q.gir s.r.o. pro OLAP [1]. V současné době jsou pomocí OLAP analýzy pokryty všechny důležité moduly IS K2 a informace v nich uložené se exportují do datového skladu. ETL proces tedy není v tomto dokumentu detailně popisován. Vývojové centrum informačního systému K2 v současné době jedná o vytvoření dolovacího modulu s Vysokou školou Báňskou v Ostravě. Tento projekt je první částí tohoto modulu s tím, že se zaměřuje pouze na jeden typ dolovací úlohy. Zbývající části modulu budou implementovány v příštím roce jako diplomové práce programátorů na oddělení vývoje informačního systému K2 z řad studentů fakulty informačních technologií na VŠB pod vedením Doc. RNDr. Jany Šarmanové, CSc.

Modul pro získávání znalostí z databází je určen především pro podporu rozhodování manažerů a ředitelů v řídící sféře podniku. Díky použití dolovacích metod je umožněn netriviální pohled na data, ze kterých se mohou získat důležité informace klíčové pro strategické rozhodování.

Tento dokument je rozdělen do následujících kapitol:

- Získávání znalostí z databází – úvod do problematiky získávání znalostí z databází, vysvětlení základních pojmů
- Informační systém K2 – popis systému, způsoby vyhodnocení
- Cíle projektu - popisuje důvod vzniku projektu
- Koncepce řešení – konceptuální návrh začlenění modulu pro dolování z dat do prostředí IS K2
- Asociační pravidla – detailní popis zvolené dolovací metody
- Implementace modulu – popis implementace modulu získávání znalostí, asociační analýzy

- Závěr – kapitola shrnuje poznatky a výsledky získané v tomto projektu, popisuje další směr vývoje produktu
- Příloha 1: Struktury číselníků parametrů
- Příloha 2: Stručná uživatelská příručka modulu získávání znalostí
- Příloha 3: Podrobný diagram tříd modulu
- Příloha 4: CD se zdrojovými texty a kódy a nápovědou k modulu

2 Získávání znalostí z databází

V dnešní době se v podnikových informačních systémech stále zvětšuje objem ukládaných dat, ovšem množství užitečných informací pro rozhodující vrstvu podniku v těchto datech je nízký. Především v řídící sféře podniku je nutné se v takovémto množství dat orientovat, vyhodnocovat potřebné informace a na jejich základě se rozhodovat. K řešení těchto úloh byly vytvořeny metody získávání znalostí z databází. Získávání znalostí z databází je obvykle definováno jako získávání skryté informace v databázi. Získávání znalostí z databází je možné aplikovat na různá uložiště dat. Typicky jsou informace dolovány z datových skladů, ovšem je možné dolovat také z relačních nebo transakčních databází, pokročilých databázových systémů (objektově orientované, prostorové, temporální, multimediální databáze) nebo WWW. Při zpracování této kapitoly jsem čerpal převážně z [2], [3] a [4].

2.1 Základní pojmy

Data – množina faktů

Databáze – určitá množina informací uložená na paměťovém médiu, obvykle se pojmem databáze označují i softwarové prostředky pro manipulaci s uloženými daty a přístup k nim. Takovýto systém se přesněji nazývá systém řízení báze dat (SŘBD).

Získávání znalost z databází (KDD – Knowledge Discovery in Databases) – je to netriviální proces získávání užitečných informací z dat. Jedná se o získávání implicitních, dosud neznámých, pochopitelných a potenciálně užitečných znalostí z databází. Proces je obvykle vícekrokový, zahrnuje přípravu dat, vyhledání vzorků, vyhodnocení znalostí,

Vzorek dat – reprezentativní vzorek dat, který zachovává platnost celého souboru faktů, které popisuje.

Užitečnost vzoru – užitečné jsou ty získané vzory, které jsou pro uživatele zajímavé. Zajímavé vzory jsou takové, které jsou platné (získaný vzorek měl být platný i pro nová data s jistým stupněm určitosti), snadně srozumitelné, použitelné (užitečné) a nové. Užitečné vzory, které přesáhnou určitý práh znalosti vzorku představují **znalost**.

Proces získávání znalostí se skládá z následujících kroků:

- **Čištění dat** – rozhodnutí, co s chybějícími daty, odstranění šumu a nekonzistence
- **Integrace dat** – sloučení dat z více zdrojů, obvykle se provádí společně s krokem čištění dat, výsledek se obvykle ukládá do datového skladu
- **Výběr dat** – výběr relevantních dat pro řešení dané analytické úlohy, v případě relačních tabulek výběr sloupců u datového skladu výběr dimenzí

- **Transformace dat** – transformace dat do konsolidované podoby, vhodné pro dolování, může jít o operace sumarizace nebo agregace. V případě použití datového skladu může transformace předcházet výběru dat a být součástí tvorby datového skladu.
- **Dolování dat** – proces aplikace zvolených metod pro řešení analytické úlohy, výsledkem je získání vzorů v datech
- **Hodnocení vzorů** – identifikace zajímavých vzorů na základě metrik užitečnosti
- **Prezentace znalostí** – prezentování získaných znalostí uživateli za použití vizualizačních a prezentačních technik

Dolovací modul – systém pro dolování informací, skládá se z několika částí:

- nastavení vstupních podmínek dolovacího modulu – grafické uživatelské rozhraní
- modul pro řešení zvolené úlohy, např. charakterizace, asociace, klasifikace, shluková analýza
- modul vyhodnocení vzorů – využívá metriky zajímavosti, rozlišuje zajímavé a nezajímavé vzory v průběhu činnosti dolovacího procesu
- prezentační vrstva – grafické uživatelské rozhraní pro prezentaci výsledků uživateli

Pro tento projekt se dolovacím modulem rozumí součást informačního systému K2, která umožňuje získávat netriviální informace z provozních dat IS K2 (budou popsány dále), skládá se z výše uvedených součástí. Grafické uživatelské rozhraní zajišťuje komunikaci uživatele s dolovacím modulem. Umožňuje nastavovat parametry dolovací úlohy, vybírat data pro dolovací úlohu a prezentuje získané znalosti uživateli.

Popisné (descriptive) dolovací úlohy – charakterizují vlastnosti zpracovávaných dat (např. statistické vlastnosti)
Prediktivní (predictive) dolovací úlohy – na základě vlastností zpracovávaných dat (data z minulosti) vytvářejí předpovědi, jaká data se budou v budoucnu objevovat (s jakou pravděpodobností)

2.2 Datové sklady

Kapitola datové sklady popisuje technologii datových skladů, která je využívána v tomto projektu. Je popsán a vysvětlen jejich vztah k informačnímu systému K2.

Data informačního systému jsou obvykle uložený v relační databázi. Pro vyhodnocení a analýzu dat však relační databáze není vhodná. Především doba potřebná pro zpracování dotazu a výpočet agregací je pro větší množství dat v porovnání s datovými sklady nesrovnatelně vyšší. Zobrazení výsledku dotazu může být uskutečněno pomocí tiskových sestav nebo exportem do jiných aplikací pomocí OLE mechanismu (např. MS Excel, MS Word). Úprava tiskových sestav či exportů

je triviální problém, ovšem časově velmi náročný. Datový sklad oproti OLTP databázi zavádí jistá vylepšení. Je optimalizován pro zadávání dotazů, především souhrnných (analýzy). Nevýhodou datového skladu ve spojení s informačním systémem K2 je neaktuálnost dat v datovém skladu. Ty se do něj exportují vždy za určité období (dávkový přístup).

Data informačního systému K2 jsou uložena v relační databázi. Informační systém K2 pracuje na různých databázových platformách:

- Pervasive SQL server
- Microsoft SQL 2000, 2005
- Oracle 10g Express Edition

Databáze IS K2 představují OLTP (On-line transactional processing) vrstvu. Tato vrstva uchovává a spravuje elementární data z různých částí podniku (např. nákup, prodej, sklad). Primárním ukazatelem efektivity této vrstvy pro uživatele je rychlost vstupu dat do systému (průchodnost systému). Pro transformaci dat z OLTP databáze do datového skladu slouží ETL procesy (Extraction, Transformation, Loading) – datové pumpy. Datovou pumpou označujeme proces, ve kterém probíhá přečtení dat z informačního systému, vyčištění a transformace do nové struktury a následně uložení do datového skladu.

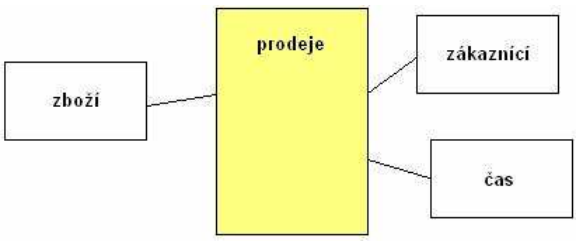
Jádrem architektury datového skladu je centrální databáze. Data v této databázi jsou dále organizována do datových tržišť, což jsou podmnožiny datového skladu. Jednotlivá datová tržiště odpovídají určitým částem podniku, jako je například nákup nebo prodej.

Výhody datového skladu se projeví až při spojení s analytickými nástroji. Ty umožňují vytváření výstupních pohledů, které slouží pro posuzování a analýzu sledovaných dat. Příkladem analytického nástroje je OLAP (Online-Analytical Processing) systém. Ten sdružuje atomická data do multidimenzionálních datových kostek, které nabízejí pohled na data z mnoha stran.

Pro modelování multidimenzionální databáze se využívají tři typy schémat:

Star

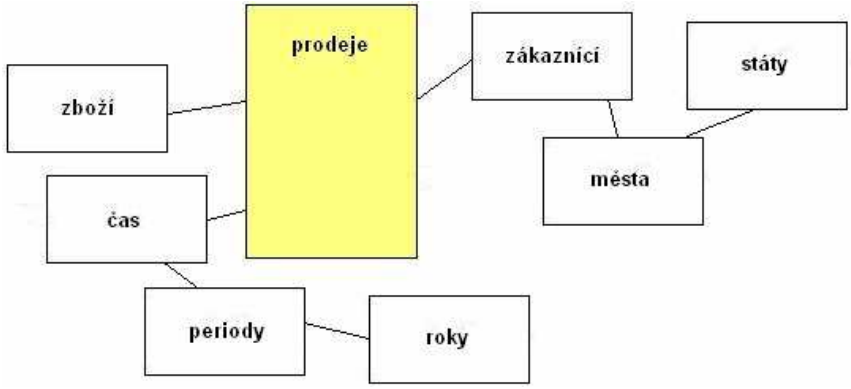
Schéma Star (hvězda, viz obr. 1) se k převodu relační databáze na multidimenzionální používá nejčastěji. Skládá se z centrální tabulky faktů s hodnotami počítaných polí a několika tabulek s osami (tabulky dimenzí). Schéma připomíná tvarem hvězdu, s centrální tabulkou uprostřed a doprovodnými tabulkami os okolo. Každá dimenze je představována právě jednou tabulkou s několika atributy.



Obr. 1 Star schéma

Snowflake

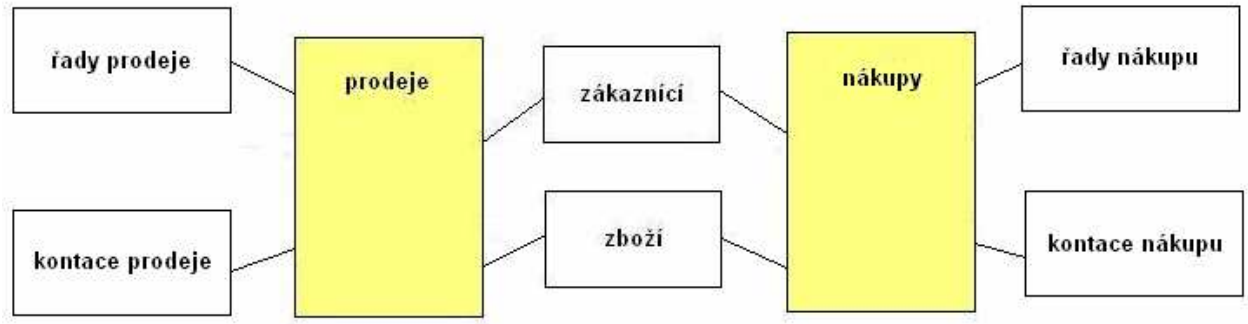
Schéma Snowflake (sněhová vločka, viz obr. 2) je hvězdicové schéma s normalizovanými tabulkami dimenzí, čímž se data rozdělují do dalších tabulek. Hlavním přínosem oproti schématu Star je snížená redundance dat a možnost lepší správy, ovšem za cenu pomalejšího přístupu v důsledku nutnosti většího počtu spojení tabulek při provádění dotazu.



Obr. 2 Snowflake schéma

Fact Constellation

V některých případech je vyžadováno propojení více tabulek počítaných polí, aby bylo možno sdílet tabulky dimenzí. Schéma vzhledem připomíná souhvězdí (Constellation, viz obr. 3) [1].



Obr. 3 Schéma Fact Constellation

2.3 Asociační analýza

Dolování asociačních pravidel (asociační analýza) se zabývá hledáním vzájemných vztahů mezi množinami dat v databázi. Hledají se dvojice tvaru atribut-hodnota, které se v datech často vyskytují společně. Nalezení těchto dvojic nad obchodními transakčními záznamy je klíčové pro obchodní rozhodování při návrhu nabídek, tvorbě katalogů nebo při rozmísťování položek v obchodě.

Typickým případem využití asociační analýzy je analýza nákupního košíku. Tento proces se zaměřuje na chování zákazníka a získává informace o tom, které položky zboží koupil zákazník společně. Např., koupí-li si zákazník klávesnici, s jakou pravděpodobností si koupí také monitor při stejném nákupu.

Typy asociačních pravidel se dělí podle:

- **typu hodnot v pravidlech:** pokud nás zajímá pouze přítomnost nebo nepřítomnost položky, pak se jedná o booleovská asociační pravidla. V případě, že asociační pravidla popisují asociace mezi kvantitativními položkami, pak se jedná o kvantitativní asociační pravidla. Příkladem může být pravidlo:

$$věk=20..25 \wedge příjem=15000..25000 \Rightarrow koupí\ (notebook)$$

- **dimenzí obsažených v pravidlech:** booleovská asociační pravidla jsou jednodimenzionální, obsahují pouze dimenzi koupí (klávesnici a monitor současně). Výše uvedené asociační pravidlo je vícedimenzionální, obsahuje celkem tři dimenze.
- **úrovní abstrakce v pravidlech:** některé metody dolování asociačních pravidel umožňují získat asociační pravidla nad více úrovněmi abstrakce. Takovýmto pravidlům říkáme víceúrovňová asociační pravidla. Příkladem mohou být následující pravidla:

$$věk=30..40 \Rightarrow koupí\ (automobil)$$

$$věk=30..40 \Rightarrow koupí\ (nákladní\ automobil)$$

2.4 Shluková analýza

Tato metoda se zaměřuje na rozdělování prvků do určitých skupin (shluků) na základě maximální podobnosti jednotlivých prvků (nejsou známy vlastnosti, podle kterých se rozděluje) uvnitř shluku a maximální odlišnosti jednotlivých shluků [3]. Podobnost objektu se posuzuje na základě hodnot jednotlivých atributů prvků a často se využívají tzv. vzdálenostní funkce. Shluková analýza se využívá například pro rozpoznávání vzorů, datovou analýzu, zpracování obrazu nebo průzkum trhu. Pro potřeby získávání znalostí z databází shluková analýza umožňuje zjistit distribuci dat a nalezení charakteristik pro jednotlivé třídy. Shlukování je možné využít v kroku předzpracování dat pro algoritmy klasifikace a predikce. Obsah této kapitoly byl čerpán z [4].

Při porovnávání jednotlivých shlukovacích metod porovnáváme jejich vlastnosti:

- **Škálovatelnost** – schopnost metody zpracovávat rozsáhlé databáze
- **Schopnost zpracovávat různé typy atributů** – schopnost metody pracovat s jinými než s numerickými hodnotami
- **Vytváření shluků různého tvaru** – mnoho metod vytváří shluky kulovitého tvaru, možnost vytvářená shluků libovolného tvaru, které lépe charakterizují analyzovaná data
- **Minimální požadavky na znalost problému při určování parametrů** – obtížné definování vstupních parametrů shlukovací metody jako je například počet požadovaných shluků
- **Filtrování šumu** – schopnost metody vyrovnat se s daty obsahujícími šum
- **Necitlivost na pořadí záznamů v databázi** – některé metody jsou při vytváření shluků citlivé na pořadí záznamů v databázi

- **Schopnost zpracovávat vysocedimenzionální data** – kvalitní algoritmy jsou schopny pracovat se záznamy o více atributech (do tří atributů je schopno analyzovat i lidské oko)
- **Schopnost shlukování na základě omezování** – nalezení shluků dat na základě daných omezení
- **Interpretovatelné a použitelné shluky** – srozumitelná reprezentace získaných shluků uživateli

Shlukovací metody pracují s daty ve formě matice (nejčastěji relační tabulka o p sloupcích a n záznamech) a musí být schopné pracovat s různými typy dat, např.: numerické, intervalové, binární, nominální (obsahují hodnoty podobné binárním ovšem s více jak dvěmi hodnotami, typicky výčet určitých typů), ordinální (obdobně jako nominální proměnné ovšem s určeným pořadím hodnot), poměrové (obsahují hodnoty měřené na nelineární stupnici) proměnné.

Existuje velké množství různých shlukovacích metod. V závislosti na činnosti jejich algoritmů je můžeme rozdělit do následujících tříd:

- **Metody založené na rozdělování** – rozdělují n objektů do k tříd, kde $k \leq n$, přičemž každá třída obsahuje alespoň 1 objekt a každý objekt patří pouze do 1 třídy. Mezi tyto metody patří například metoda k-means.
- **Hierarchické metody** – tyto metody vytvářejí hierarchický rozklad dané množiny objektů, přičemž vzniká strom shluků. Mezi tyto metody patří například metody AGNES a DIANA.
- **Metody založené na hustotě** – tyto metody považují za shluky oblasti s velkou hustotou objektů v prostoru dat, které jsou od sebe odděleny oblastmi s malou hustotou objektů. Objekty vyskytující se v prostoru s malou hustotou objektů jsou považovány za šum. Tyto metody mají výhodu v tom, že dokáží vytvářet shluky různých tvarů a dobře se vypořádávají s šumem a odlehlými hodnotami. Mezi tyto metody patří například metody DBSCAN a DENCLUE.
- **Metody založené na mřížce** – využívají víceúrovňové mřížkové struktury. Prostor objektů rozdělují na určitý počet buněk, které tvoří mřížku. Všechny operace shlukování probíhají nad touto mřížkovou strukturou. Mezi hlavní výhody patří rychlá doba zpracování datové množiny, která je nezávislá na objemu dat, pouze na počtu buněk mřížkové struktury. Mezi tyto metody patří například metody STING a WaveCluster.
- **Metody založené na modelech** – tyto metody se snaží optimalizovat shodu mezi datovou množinou a nějakým matematickým modelem.Tyto metody jsou nejčastěji založeny na tom, že data jsou generována na základě nějaké složené pravděpodobnostní distribuční funkce. Mezi tyto metody patří například metody Expectation-Maximization, konceptuální shlukování nebo metody neuronových sítí.

- **Metody pro shlukování vysocedimenzionálních dat** – tyto metody se snaží minimalizovat problémy se zpracováním dat o vysokém počtu dimenzí, u kterých je obtížné a někdy i nemožné najít shluky, protože pouze některé dimenze jsou pro zařazení do shluků relevantní (ostatní produkují šum). Tyto metody řeší výše uvedené problémy buď transformací dat do prostoru s nižším počtem dimenzí při zachování relativní vzdálenosti mezi objekty (sloučením několika parametrů do jednoho, což ovšem komplikuje následnou interpretaci shluků) nebo pomocí snížení počtu atributů (výběr relevantních atributů, učení s učitelem). Mezi tyto metody patří například metody C1IQUE nebo PROCLUS.

2.5 Klasifikace a predikce

Metoda se zaměřuje na určování příslušnosti dat do předem známých skupin nebo tříd. Typickým příkladem podobné analýzy je práce bankovního experta, který se na základě informací o zákaznících (např. věk, příjem, profese, platební morálka) rozhoduje, zda dané osobě poskytne úvěr. Na základě předchozích případů určí, které z vlastností jsou pro daný problém důležité a s určitou pravděpodobností poté vyhodnotí riziko poskytnutí úvěru novému zákazníkovi (pouze na základě znalosti některých vlastností zákazníka). Pojem klasifikace vyjadřuje proces, kterým se daný objekt zařadí do jisté třídy na základě jeho vlastností, pojem predikce pak předpověď jisté hodnoty (obecně spojitého charakteru) pro daný objekt na základě jeho vlastností.

Ve fázi klasifikace je nejprve nutné zjistit tzv. klasifikační pravidla, pomocí kterých se s jistou přesností objekt klasifikuje do dané třídy. Tato fáze probíhá na tzv. trénovací množině dat, u které známe správně zařazení do tříd. Poté se získané pravidla testují pomocí testovací množiny dat, u které taktéž známe správné zařazení do tříd (ovšem algoritmus o něm neví) a procentuálně se určí na kolik byl algoritmus při klasifikaci úspěšný. Velikost této procentuální hodnoty určuje úspěšnost a použitelnost testovaného klasifikačního pravidla do budoucna.

Klasifikační metody lze porovnávat na základě následujících kritérií:

- **Přesnost předpovědi** – procentuální hodnota správné klasifikace nových dat (data která nabyla v testovací množině)
- **Rychlost** – výpočetní složitost pro vygenerování a používání klasifikačních pravidel
- **Robustnost** – schopnost vytvoření správného modelu ze zašuměných dat
- **Stabilita** – nezávislost vytvoření modelu na objemu dat
- **Interpretovatelnost** – složitost modelu pro pochopení

Metody klasifikace mohou pracovat na základě rozhodovacího stromu, což je graf stromové struktury, kde každý vnitřní uzel reprezentuje test hodnoty jistého atributu a listy reprezentují třídu, do které byl daný objekt klasifikován. Mimo tvorby klasifikátoru pomocí rozhodovacího stromu lze klasifikaci založit také na statistice. Metoda využívající statistiku se nazývá Bayesovská klasifikace.

Pracuje na principu určení pravděpodobnosti, s jakou daný objekt patří do určitých tříd. Objekt je zařazen do té třídy, do jaké s největší pravděpodobností patří. Dalším modelem pro klasifikaci mohou být neuronové sítě (Backpropagation), kde se na vstupní vrstvu neuronů přivede klasifikovaný objekt a na výstupní vrstvě se získá zjištěná třída.

Ve fázi predikce se nejčastěji využívá jednoduchá nebo násobná lineární regrese. Nelineární úlohy lze na lineární často transformovat. Při použití jednoduché lineární regrese se nejprve pomocí metody nejmenších čtverců zjistí rovnice přímky, která aproximuje data. Podklady této kapitoly byly čerpány z [4].

3 Informační systém K2

Informační systém K2 vyvinula společnost Q.gir s.r.o. v roce 1992 za použití programovacího jazyka Clarion. Původní verze pro DOS byla v roce 1999 implementována pro operační systém Windows v programovacím jazyku Delphi. Největší důraz je v něm kladen na přesnou znalost obchodní a finanční situace firmy. Dále se klade vysoký důraz na dobrou organizaci práce, aby nedocházelo ke ztrátám pracovního a materiálního potenciálu firmy. To vytváří silný tlak na pracovníky v řídící sféře. Ti jsou nuceni vyhodnocovat, zpracovávat a následně i využívat data proudící ve firmě. Nástrojem pro tuto činnost se stává software, který umožňuje uživateli rychlou, podrobnou a přesnou evidenci dat. Program K2 je informační systém. To znamená, že veškeré údaje, data a informace jsou přímo vkládány do programu, všechny doklady se pořizují prostřednictvím tohoto programu a následně je zajištěna také vazba na všechny strany, kde je zapotřebí vkládané údaje využívat. Program K2 je schopen zabezpečit zpracování dat a zaručit jejich správnost. Předností IS K2 je schopnost provádět rozsáhlé kontroly konzistence vstupních dat, kdy výpočet vyhodnotí, ve kterém místě došlo k narušení spojitosti dat a dále je schopen provádět přepočty veškerých výstupních údajů.

K2 je systém, který zachycuje oběh prvotních dokladů ve společnosti na základě jejich skutečného oběhu a fyzického stavu. Z toho vyplývá, že K2 není účetní program, ale informační systém, jehož cílem je evidovat probíhající procesy. Informační systém K2 podporuje pesimistické holdování a transakce s možností zanoření.

IS K2 spadá do třetí generace databázových manažerských systémů (dále jen DBMS) a jako takový splňuje následující zásady a skupiny tvrzení:

Zásady

DBMS třetí generace poskytují služby ve třech oblastech: správa dat, objektů a znalostí. Jsou otevřeny jiným subsystémům. Obsahují prostředky pro podporu rozhodování, operace se vzdálenými daty a distribuované zpracování.

Skupiny tvrzení

Třetí generace DBMS obsahuje systém bohatých datových typů. Podporuje dědičnost, funkce (metody, a databázové procedury) a zapouzdření. Správa UID a vyrovnávací paměti jsou na fyzické úrovní a ne v datovém modulu. IS K2 nepodporuje programátorsky specifikovaný přístup k datům. IS K2 je multijazyčný. V prostředí IS K2 je úzká shoda mezi typovým systémem a implementačním jazykem. SQL dotazy a jejich výsledky jsou nejnižší komunikační úrovní mezi klientem a serverem.

3.1 Současný stav vyhodnocení informací

Během zpracování dat v informační systému je podstatné jak jejich správné zadávání, tak také především jejich vyhodnocení. K vyhodnocení výsledků je potřeba zvolit správný nástroj. Typy nástrojů pro vyhodnocení dat je možno rozdělit do následujících úrovní:

- **Nástroj pro rozhodovaní na nižší úrovni**
  - objednávky zboží, plnění norem pracovníků, odměny a prémie pracovníků
  - pohledy IS na data (formuláře s předem připravenými analýzami), tiskové sestavy, objednávkové automaty
- **Nástroj pro kontrolu**
  - inventura skladu, kontrola správně zadaných dat
  - tiskové sestavy, automatizované přepočty (fyzické, účetní), skladové kontrolní mechanizmy (automatizovaný sběr dat pomocí čteček čárových kódů)
- **Nástroj pro manažerské rozhodování**
  - strategická rozhodnutí ve společnosti (Vyplatí se výroba výrobku? Je poptávka po výrobku dostačující?)
  - statistiky, přehledové grafy, OLAP analýza

Forma výsledku a jeho podrobnost velice ovlivňuje smysluplnost výsledku a do jisté míry ovlivňuje také použitelnost tohoto výsledku.

3.2 Skriptovací jazyk K2

Skriptovací jazyk K2 (dále jen K2S) je postaven na bázi jazyka Pascal a umožňuje vytvářet programy v systému K2. Tyto programy mohou číst i zapisovat data, vytvářet doklady, spouštět tiskové sestavy, či jinak modifikovat celkové chování informačního systému K2. Zdrojové soubory programu je možno psát v libovolném textovém editoru, nejlépe však v editoru skriptu, který je součástí K2 a má zabudovány ladící funkce. Obsahuje také šablony kódu, kontextovou nápovědu (skript asistent) a různé způsoby průběžného vyhodnocování během ladění (watchlist). Základní programové konstrukce, datové typy a dialogy odpovídají standardnímu Pascalu. Rozšířením K2 skriptu oproti Pascalu je zavedení objektů (obdobně jako v Delphi). Pro přístup a práci s daty jsou v programovacím jazyku K2 skript implementovány různé třídy, jejichž funkci zde nyní popíši.

TxFile – předek všech dále uvedených tříd pro přístup datům. Třída obsahuje metody pro práci s daty (např. vkládání a rušení prvku) a pro pohyb po datech. Pro posun v datech se využívá stabilní kurzor, který je nastavitelný pomocí metod třídy. Uspořádání kolekce je dáno nastavením určitého klíče a hodnot jeho atributů.

Memfile – potomek třídy TxFile. Jedná se o třídu, která obsahuje kolekci záznamů programátorem definovaného typu a tím vytváří obdobu relační tabulky, která je však pouze v paměti a neukládá se

do databáze. Po ukončení skriptu je její obsah zahozen. Na datech memfile je vždy nutné definovat unikátní index.

AdoFile – tato třída je potomkem TxFile obdobně jako třída Memfile, ovšem s tím rozdílem, že kolekce záznamů třídy je mapována na konkrétní tabulku v databázi. Atributy této třídy jsou tedy sloupce databázové tabulky. Od této třídy jsou zděděny všechny třídy v K2 skriptu pro přístup k datům (např. třídy "Zbozi" nebo "Zakazky"). Konvence K2 skriptu pojmenovává všechny třídy pro přístup k datům tabulek K2 podle názvu těchto tabulek. Na datech je možné definovat index.

DosFile – tato třída je potomkem třídy TxFile. Slouží pro práci s textovými soubory. Její atributy jsou v textové podobě ukládány do definovaného souboru na disku (jeden záznam této reprezentuje jeden řádek). Tato třída neumožňuje definici indexu na datech.

tDataM (datový modul) – tato třída je potomek třídy TxFile. Kolekce záznamů této třídy je mapována na tabulku databáze K2 stejně jako v případě třídy TxFile, ovšem s tím rozdílem, že datové moduly obsahují také všechny podřízené tabulky (v podobě tzv. podřízeného datového modulu), na které má hlavní tabulka třídy vazbu (především vazbu 1..N). Vazby mezi datovými moduly (vztahy) jsou dány strukturou databáze K2. Datový modul je v informačním systému K2 nositelem persistence dat. Příklad použití datového modulu zakázek je následující:

- Nastavení indexu: Zak.AktIndexDM := 0;
- Nastavení na první záznam: Zak.SetFirstDM;
- Pohyb po hlavičkách: Zak.DonextDM;
- Čtení atributu zakázky: idDodatvatele := Zak.CDo;
- Nastavení na první položku zakázky: Zak.NakPro.SetFirstDM;
- Pohyb po položkách zakázky: Zak.NakPro.DoNextDM;
- Uložení zakázky: Zak.UlozDM;

3.3 Moduly informačního systému K2

Informační systém K2 je rozdělen do několika modulů, které sdružují logicky související části provozního systému. Tyto moduly obsahují práci s doklady, které jsou nutné pro funkčnost určitého modulu. Z důvodu omezeného rozsahu tohoto projektu se zaměřím pouze na dva základní moduly a to konkrétně na nákup a prodej. Informace v následujících kapitolách byly čerpány z [6] a [7].

3.3.1 Nákup

Modul Nákup slouží k řízení nákupu zboží, surovin a služeb od dodavatelů. Hlavním dokladem tohoto modulu je tzv. kniha objednávek. Na tomto dokladu se zadává především zvolený dodavatel a objednávané položky. Je zde znázorněn stav zpracování objednávky (plánovaný nebo potvrzený termín dodání, existence příjemky, stav naskladnění zboží, existence přijaté faktury atd.) po

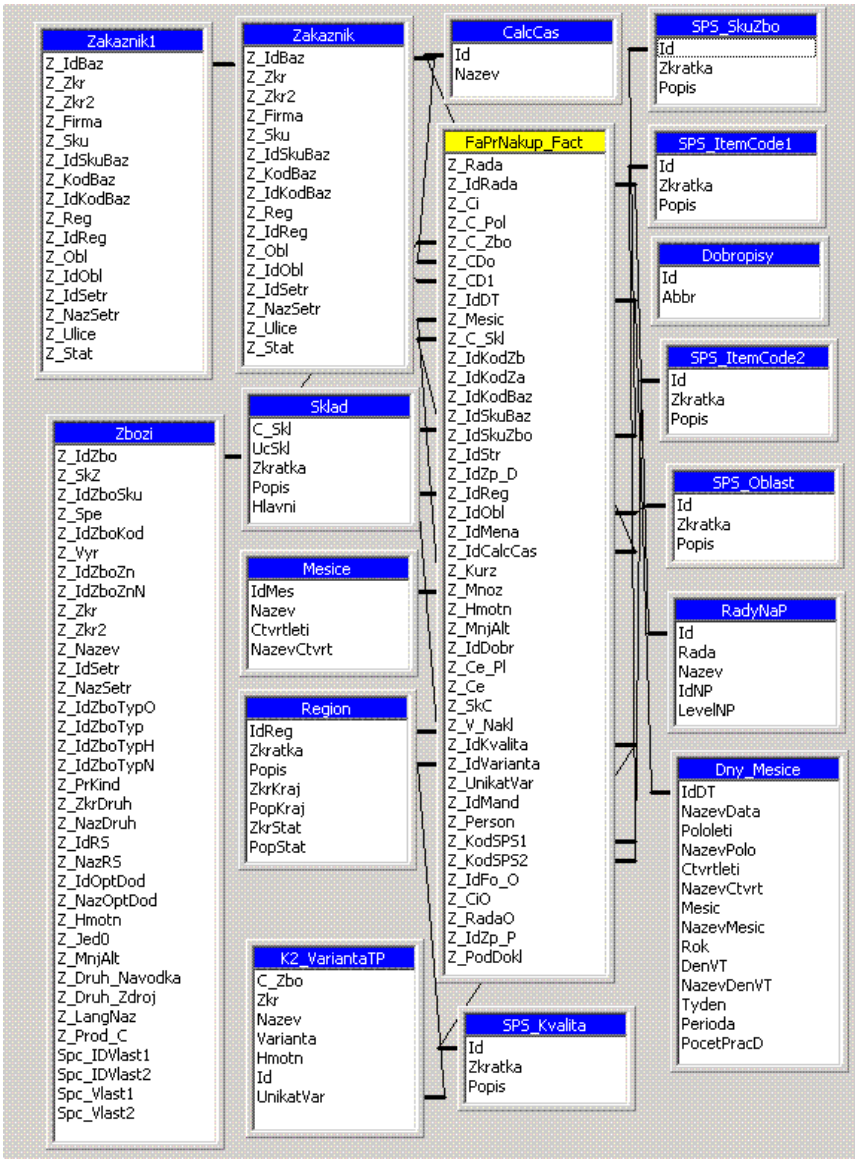
jednotlivých položkách. Toto je technicky zabezpečeno tak, že každá položka objednávky na sobě nese vazbu na případnou příjemku, fakturu atd. Díky tomuto návrhu je v dokladech nákupu zabezpečena silná kontrola konzistence dokladů a není možné například zadat fakturu na položky, která nebyla v objednávce a naopak nelze na některou objednávanou položku neúmyslně zapomenout.

Zboží lze objednávat buď cyklicky, tzv. na sklad, nebo je možné objednávat adresně na konkrétní zakázky s vazbou na okamžitý prodej zákazníkovi. U každé nakupované položky je možné evidovat termíny založení a odeslání objednávky a předpokládané a skutečné termíny dodání. Evidované termíny rozlišují množství a nákupní ceny, u položek je možné evidovat vlastnosti a jakosti nakupovaného zboží. To je velmi významné při nakupování materiálů pro výrobu, protože kromě jasného určení, z čeho byl daný kus výrobku vyroben, vkládají nákupy do ocenění výrobku skutečné ceny vstupů. U nákupů je dále možné evidovat vedlejší pořizovací náklady a ty následně rozpouštět do jednotlivých položek, případně celého dokladu.

Modul řeší vystavování takzvaných traťových dodávek, při kterých se zboží dodává přímo od výrobce zákazníkovi a vůbec fyzicky nevstupuje na sklad prodejce. Modul pracuje s čárovými kódy, kódy zakázek, šaržemi, sériovými čísly, umístěním (lokace na skladě) a dalšími specifickými označeními (např. kódy zákazníka apod.). Modul umožňuje definovat optimálního dodavatele pro jednotlivé nakupované položky. U každé nakupované položky je poté možné evidovat hodnocení dodavatele a to především z pohledu rychlosti, spolehlivosti, kvality, jakosti, dodržení ceny, množství atd. Zvláště hodnocení dodavatele je dobrý potenciální zdroj dat například pro predikci dodržení ceny, charakterizaci dodavatelů, kteří nedodržují některé kriterium hodnocení (shlukování). Metodu analýzy nákupního košíku by v tomto případě mohla být použita (za předpokladu, že bychom za nákupní košík označili například objednávky z určitého časového období), ovšem její výsledky by bylo možné použít pouze k optimalizaci nákupů vyhledáním vhodných optimálních dodavatelů.

Datový trh Nákup

Pro potřeby získávání znalostí z databází je vhodné využít datový sklad (viz. kapitola 2.2). Data do tohoto datového skladu jsou exportována pomocí exportních skriptů původně vytvořených pro potřeby OLAP analýzy (více viz. bakalářské práce [1]). Jak již bylo uvedeno výše, primárním nositelem informace o uskutečněných nebo plánovaných nákupech jsou položky objednávek. Tyto položky tvoří základ faktové tabulky, která je jádrem hvězdicové struktury datového trhu Nákup. Tabulky dimenzí jsou tvořeny číselníky, které jsou na tabulku faktů navázány (typicky zboží, zákazníci, regiony atd.). V případě zvolení tohoto modulu (nákup) by algoritmy dolování znalostí využívaly data uložená v této části datového skladu. Schéma datového trhu viz obr. 4.



Obr. 4 Hvězdicové schéma datového trhu Nákup

3.3.2 Prodej

Modul prodej slouží k realizaci obchodních případů zákazníků. Hlavním dokladem tohoto modulu je tzv. kniha zakázek. Na tomto dokladu se zadává především zvolený odběratel (pomocí metody shlukování je možné zjišťovat geografické rozmístění odběratelů pro zacílení reklamy) a prodávané (nabízené) položky (toto je typický příklad analýzy nákupního košíku). Je zde znázorněn stav zpracování zakázky (plánovaný nebo potvrzený termín dodání, existence výdejky, stav vyskladnění zboží, existence dodacího listu, existence vydané faktury atd.) po jednotlivých položkách. Toto je technicky zabezpečeno tak, že každá položka zakázky na sobě nese vazbu na případné další doklady prodeje, obdobně jako u položek nákupu na doklady nákupu (viz. předchozí kapitola).

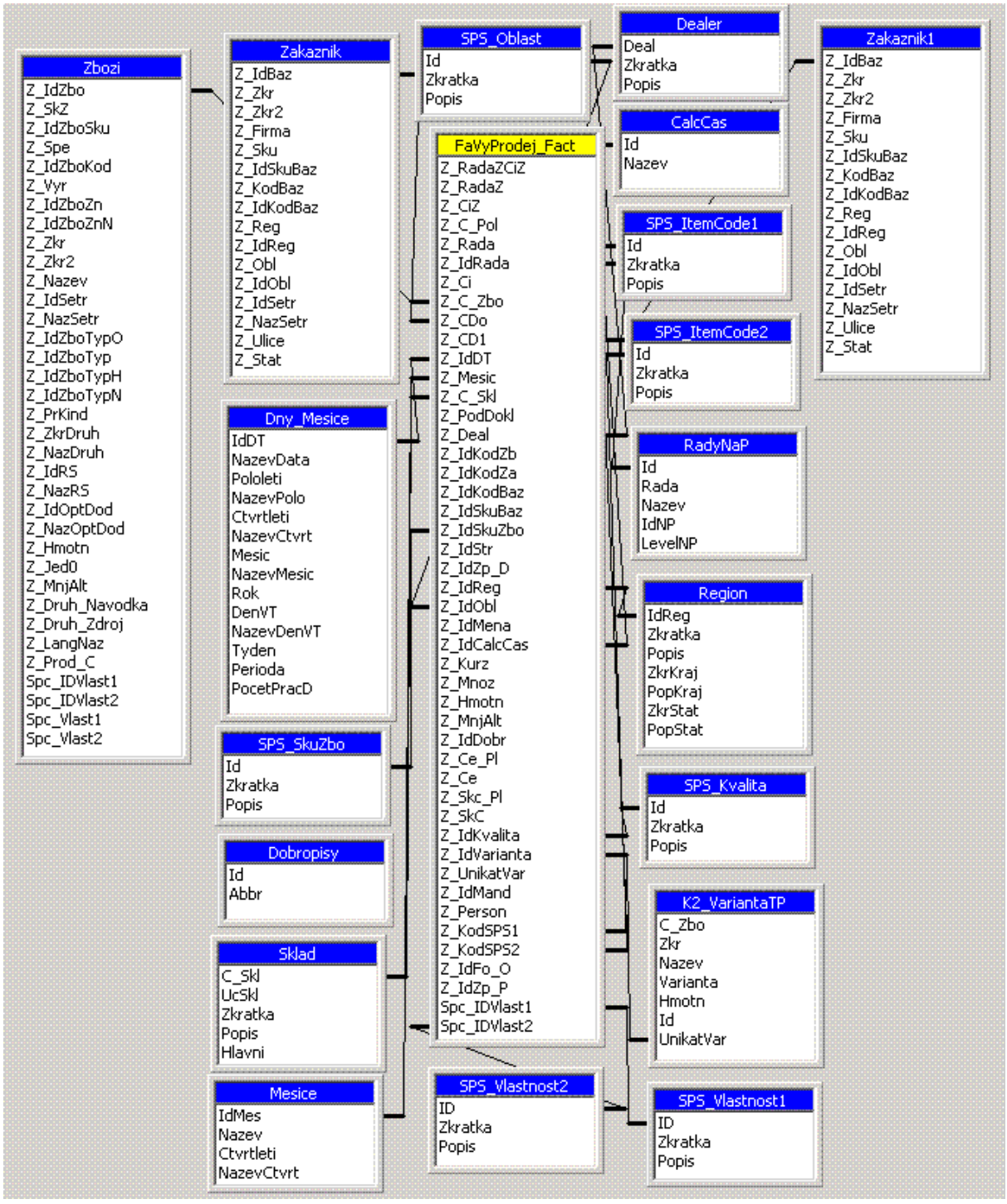
U každé prodávané položky je možné evidovat termíny založení a odeslání nabídky a předpokládané a skutečné termíny dodání (úloha predikce: realizace nabídky). Evidované termíny rozlišují množství a prodejní ceny, u položek je možné evidovat vlastnosti a jakosti prodávaného zboží (důležité zvláště pro vyráběné položky). Modul umožňuje evidovat změny prodejních cen oproti standardně definované prodejní ceně. Prodejní cenu je možné ovlivňovat přímo jejím nastavením, nebo použitím slevy, případně přirážky (pevná, procentuální). Důležitou vlastností položky je příznak rezervace (který vytváří rezervační list), který signalizuje požadavek na nákup (výrobu) položky obchodníkovi (výrobnímu manažerovi).

Modul pracuje s čárovými kódy, kódy zakázek, šaržemi, sériovými čísly, umístěním (lokace na skladě) a dalšími specifickými označeními (např. kódy odběratelů apod.).

Modul implementuje systém potvrzování dokladů, který jednoznačně identifikuje realizované prodeje a nabídky (nerealizované prodeje). Díky tomuto systému lze jednoznačně identifikovat zákazníky, kteří neustále poptávají, a nikdy nerealizují nabídková řízení (úloha identifikace vlastností takovýchto zákazníků pomocí metody shlukování).

Datový trh Prodej

Obdobně jako u modulu Nákupu (viz. předchozí kapitola) jsou data Prodeje exportována do datového skladu pomocí exportních skriptů pro OLAP analýzu. Primárním nositelem informace o nabídkách a zakázkách (uskutečněných nabídkách) je analogicky položka zakázky, která tvoří základ faktové tabulky datového trhu Prodej. Tabulky dimenzí jsou tvořeny číselníky, které jsou na tabulku faktů navázány (typicky zboží, zákazníci atd.). V případě zvolení tohoto modulu (prodej) pro dolování dat by algoritmy dolování znalostí využívaly data uložená v této části datového skladu. Schéma datového trhu viz obr. 5.



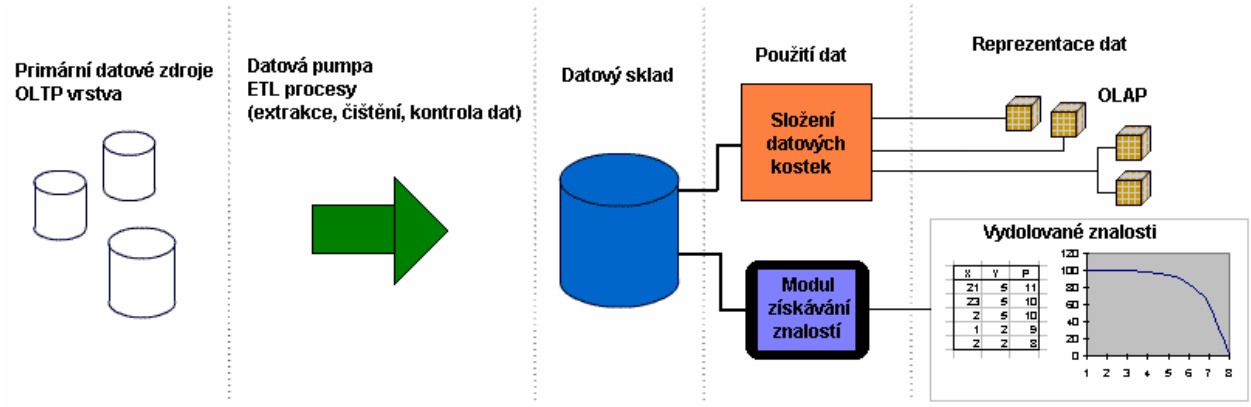
Obr. 5 Hvězdicové schéma datového trhu Prodej

4 Cíle projektu

Cílem projektu je získávání netriviálních znalostí z dat informačního systému K2 pro účely rozhodování, řízení a plánování. V předchozích kapitolách byly představeny současné metody pro vyhodnocení dat. S použitím těchto metod lze získat triviální informace o datech společnosti, ovšem již není možné získat informace o skrytých, na první pohled nezřejmých, závislostech mezi nimi. K těmto účelům slouží právě metody dolování znalostí z databází. Během konzultací se zákazníky jsem obdržel následující typy analytických úloh:

- Zobrazení geografického rozmístění zákazníků na mapě (konkrétně grafické znázornění oblastí s různými objemy prodejů v jednotlivých městech na mapě), úloha shlukování
- Důvody odmítnutí nabídek (konkrétně charakterizace nabídek zákazníkům, u kterých nebyla realizována nabídka v podobě smlouvy), úloha charakterizace a predikce
- Analýza nákupního košíku nad položkami prodeje (konkrétně příprava nabídek, dárkových balení, optimalizace rozmístění položek v obchodě)

5 Koncepce řešení - architektura systému



Obr. 6 Architektura dolovacího systému

Data z informačního systému jsou získávána pomocí datové pumpy, která je následně uloží do speciální databáze – datového skladu, který pak umožňuje efektivní a rychlou analýzu dat, bez nutnosti procházení OLTP databáze v době dotazu. Možné použití dat je různé, např. vytvoření datových kostek nebo zpracování pomocí modulu pro získávání znalostí. Prezentační vrstva zajišťuje komunikaci uživatele se systémem. Uvedená architektura viz obr. 6.

Na základě předložených dotazů a po konzultaci s vedoucím projektu ze strany firmy i školy jsem se rozhodl pro implementaci asociační analýzy nad daty datového trhu Prodej pomocí algoritmu Apriori.

Funkční a nefunkční požadavky

Funkční a nefunkční požadavky tvoří základní způsob specifikace výsledného systému. Jedná se o vlastnosti a funkce, které musí systém realizovat a podporovat. Při zpracování této kapitoly jsem vycházel z [1].

Funkční požadavky:

- Výsledný modul pro získáván znalostí z databází musí umožňovat analýzu vybraných dat založenou na zvolené dolovací úloze.
- Údaje pro vyhodnocení musí být získávány z datového skladu.
- Data budou do datového skladu exportována pomocí datové pumpy, která je součástí exportních skriptů modulu IS K2 pro OLAP analýzu.

Nefunkční požadavky: Mezi základní nefunkční požadavky na výsledný dolovací modul patří:

- Snadná obsluha dolovacího modulu: krátké zaškolení uživatelů, intuitivní nastavení parametrů dolovací úlohy, přehledné zobrazení výsledků, možnost uložení výsledků
- Univerzálnost, znuvupoužitelnost, rozšiřitelnost: snadné definování požadavků, možnost uložení nastavení
- Bezpečnost: analýza dat datového skladu, možnost opětovného vygenerování, neexistuje riziko poškození provozních dat společnosti, možnost nastavit práva jednotlivých uživatelů na přístup do dolovacího modulu
- Rychlost a výkonnost: zpracování dat datového skladu dolovacím algoritmem musí proběhnout v omezeném čase (v závislosti na objemu dat řádově několik desítek hodin), dolovací modul musí být schopen zpracovat vstupní data nezávisle na jejich objemu (s možností využití omezujících – filtrujících podmínek), současný přístup do dolovacího modulu musí být umožněn několika uživatelům informačního systému K2 (závisí na nastavení uživatelských práv)
- Validita výstupních dat: systém musí poskytovat korektní výsledky, výsledky musí být platné v době zpracování dat

6 Asociační pravidla

Tato kapitola se detailněji zaměřuje na problematiku asociační analýzy a rozšiřuje úvodní kapitolu 2.3. Při jejím zpracování jsem čerpal především z [3] a [4].

Jak již bylo uvedeno výše, dolování asociačních pravidel se zaměřuje na hledání vzájemných vztahů mezi množinami dat v databázi. Tyto vztahy nazýváme asociační pravidla. Jsou to implikace tvaru $A \Rightarrow B$, což odpovídá tomu, že pokud je v dané množině hodnot obsažena položka *A*, je v ní s velkou pravděpodobností i položka *B*. Typické použití asociační analýzy je v oblasti obchodu, konkrétně při analýze nákupního košíku. Implikaci $A \Rightarrow B$ lze tedy popsat jako: Koupí-li si zákazník položku *A*, koupí si s velkou pravděpodobností i položku *B*.

Definice:

Převzato z [4]. Nechť $I=\{ i_1, i_2, i_3,...\}$ je množina položek, *D* je množina transakcí, kde každá transakce *T* je množina položek taková, že $T \subseteq I$. Ke každé transakci přísluší unikátní identifikátor *TID*. Nechť *X* je množina položek. Říkáme, že transakce T obsahuje *X* tehdy a jen tehdy, pokud $X \subseteq T$.

Asociační pravidlo je implikace tvaru $A \Rightarrow B$, kde $A \subset T$, $B \subset T$ a $A \cap B = \varnothing$.

Pravidlo $A \Rightarrow B$ má **podporu (support) s** v množině transakcí *D*, jestliže s% transakcí v *D* obsahuje množinu položek $A \cup B$.

Pravidlo $A \Rightarrow B$ má v množině transakcí **spolehlivostí (confidence) c**, jestliže c% transakcí, které obsahují *A* obsahuje také *B*.

S využitím pravděpodobnosti lze tedy napsat:

$$s(A \Rightarrow B) = P(A \cup B) \quad (1)$$

$$c(A \Rightarrow B) = P(B|A) \quad (2)$$

Pomocí těchto dvou parametrů (podpory a spolehlivosti) lze určit, jak statisticky významné pravidlo jsme z databáze získali, tj. jak je pro nás zajímavé. Obyčejně volíme nějakou spodní mez minimální podpory a minimální spolehlivosti, pomocí kterých se vylučují ta pravidla, která pro nás nejsou tak zajímavá [3].

Získávání asociačních pravidel probíhá v následujících krocích:

1) Získání frekventovaných množin (množin položek), které splňují podmínku minimální podpory.
2) Získání silných asociačních pravidel z frekventovaných množin. Tato pravidla musí splňovat podmínku minimální podpory a minimální spolehlivosti.

Množina obsahující právě k položek nazýváme k-množinou. Frekvence výskytu množiny je počet transakcí, které ji obsahují [4].

6.1 Jednoúrovňová booleovská asociační pravidla

Nejjednodušší typ analýzy je dolování jednoúrovňových asociačních pravidel (popis jednoúrovňových asociačních pravidel viz kapitola 3.3) a nejpoužívanější algoritmus pro získávání těchto pravidel je algoritmus Apriori nebo jeho modifikace.

6.1.1 Algoritmus Apriori

Apriori je algoritmus pro získávání frekventovaných množin. Název algoritmu je založen na faktu, že algoritmus využívá předchozí znalost o dříve vygenerovaných frekventovaných množinách. V každé iteraci získané frekventované k-množiny jsou použity pro generování (k+1)-množin. Každá iterace ta vyžaduje průchod databází. K vyšší efektivitě se využívá tzv. Apriori vlastnost. Ta říká, že každá podmnožina frekventované množiny musí být také frekventovaná. To vychází z faktu, že přidání prvku k množině nemůže způsobit, že by její podpora vzrostla. Proces aplikace Apriori vlastnosti se skládá ze dvou kroků:

- Spojovací krok: K nalezení L_k (všech frekventovaných k-množin) kandidáti na frekventované množiny (C_k) jsou vygenerováni spojením množin L_{k-1}. Nechť l_1 a l_2 jsou množiny L_{k-1}. Algoritmus předpokládá, že položky v množině jsou lexikograficky seřazeny. Ke spojení dvou (k-1)-množin dojde, pokud jejich prvních k-2 prvků je shodných. Výsledná množina vzniklá spojením l_1 a l_2 bude mít prvky $l_1[1], l_1[2], ..., l_1[k-1], l_2[k-1]$, kde $l_1[i]$ je i-tý prvek množiny l_1 a musí platit, že $l_1[k-1] < l_2[k-1]$.
- Vylučovací krok: C_k je nadmnožinou L_k, tzn., že její prvky buď jsou nebo nejsou frekventované. Zjišťování výskytu každého kandidáta v databázi je velmi neefektivní, proto využíváme Apriori vlastnost, Žádná (k-1)-množina která není frekventovaná, nemůže být součástí frekventované množiny. Tedy v případě, že některá (k-1)-podmnožina není frekventovaná, nemůže být frekventovaná ani kandidátní k-množina a můžeme ji odstranit z C_k. Toto vyhledávání podmnožin může být zrychleno využitím hašovacího stromu pro frekventované množiny.

6.1.2 Generování asociačních pravidel z frekvent. množin

Toto generování je založeno na využití rovnice pro výpočet spolehlivosti:

$$conf(A \Rightarrow B) = P(B|A) = \frac{s(A \cup B)}{s(A)} \quad (3)$$

Generování asociačních pravidel pak postupuje podle následujících kroků:

- Pro každou frekventovanou množinu *l* generuj všechny její neprázdné podmnožiny.
- Pro každou podmnožinu *s* vygeneruj pravidlo $s \Rightarrow (l - s)$ a podle rovnice vypočti jeho spolehlivost. Pokud je jeho spolehlivost vyšší než minimální, pak je pravidlo silné. Protože jsou pravidla generována z frekventovaných množin, tak pro ně automaticky platí podmínka minimální podpory.

6.2 Víceúrovňová asociační pravidla

V některých případech je složité najít zajímavé asociace v datech, zvláště pak v případě, když je velikost množiny položek příliš veliká. Typickým příkladem je asociační pravidlo typu: koupí-li si zákazník pevný disk Seagate ST3400620NS, koupí si i myš Microsoft Wheel Mouse Optical. Toto pravidlo bude mít pravděpodobně malou podporu, ovšem pokud vneseme informaci o tom, že všechny typy pevných disků patří do množiny HDD a všechny typy myší do množiny MOUSE, pak asociační pravidlo typu: koupí-li si zákazník HDD, koupí si i MOUSE bude mít podporu jistě větší bez významné ztráty informace. Takovéto zvyšování abstrakce je definováno jako obecný strom, jehož listy jsou jednotlivé položky a uzly tohoto stromu jsou pak jejich generalizace (zobecnění).

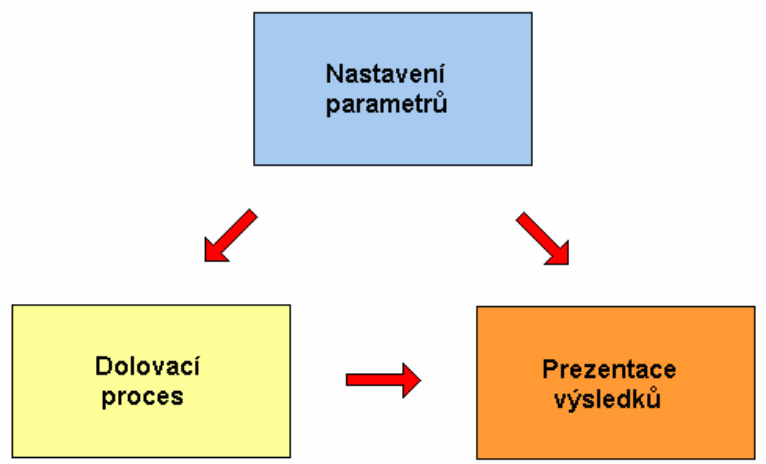
Při použití abstrakce platí pro definici asociačních pravidel jistá omezení. Tvar asociačního pravidla $A \Rightarrow B$ zůstává, ovšem platí, že *A* a *B* jsou disjunktní množiny položek (žádná položka z A není obsažena v množině B). K získávání víceúrovňových asociačních pravidel se používá algoritmus Apriori a to tak, že se prochází hierarchie stromu položek shora dolů a na každé úrovni se aplikuje algoritmus Apriori. Podpora a spolehlivost se počítá stejně jako u jednoúrovňových pravidel, ovšem s tím rozdílem, že za výskyt položky v transakci se považuje buď její konkrétní výskyt nebo výskyt předchůdce položky v hierarchii stromu. Tímto způsobem získaná pravidla posuzujeme na redundantnost. Redundantní asociační pravidla se vyznačují tím, že je možné je přímo odvodit z pravidla, které již bylo získáno na vyšší úrovni. Získáme-li například dvě asociační pravidla:

- koupí počítač $\Rightarrow$ koupí myš
- koupí počítač Acer $\Rightarrow$ koupí myš,

Pak je druhé asociační pravidlo redundantní, nepřináší žádné nové poznatky a má nižší podporu než první pravidlo. Pravidlo R1 nazýváme předchůdcem pravidla R2, jestliže R1 dostaneme nahrazením položek R2 jejich předchůdci v konceptuální hierarchii. Pravidlo je redundantní, má-li očekávané hodnoty podpory a spolehlivosti s ohledem na předchůdce.

7 Implementace modulu

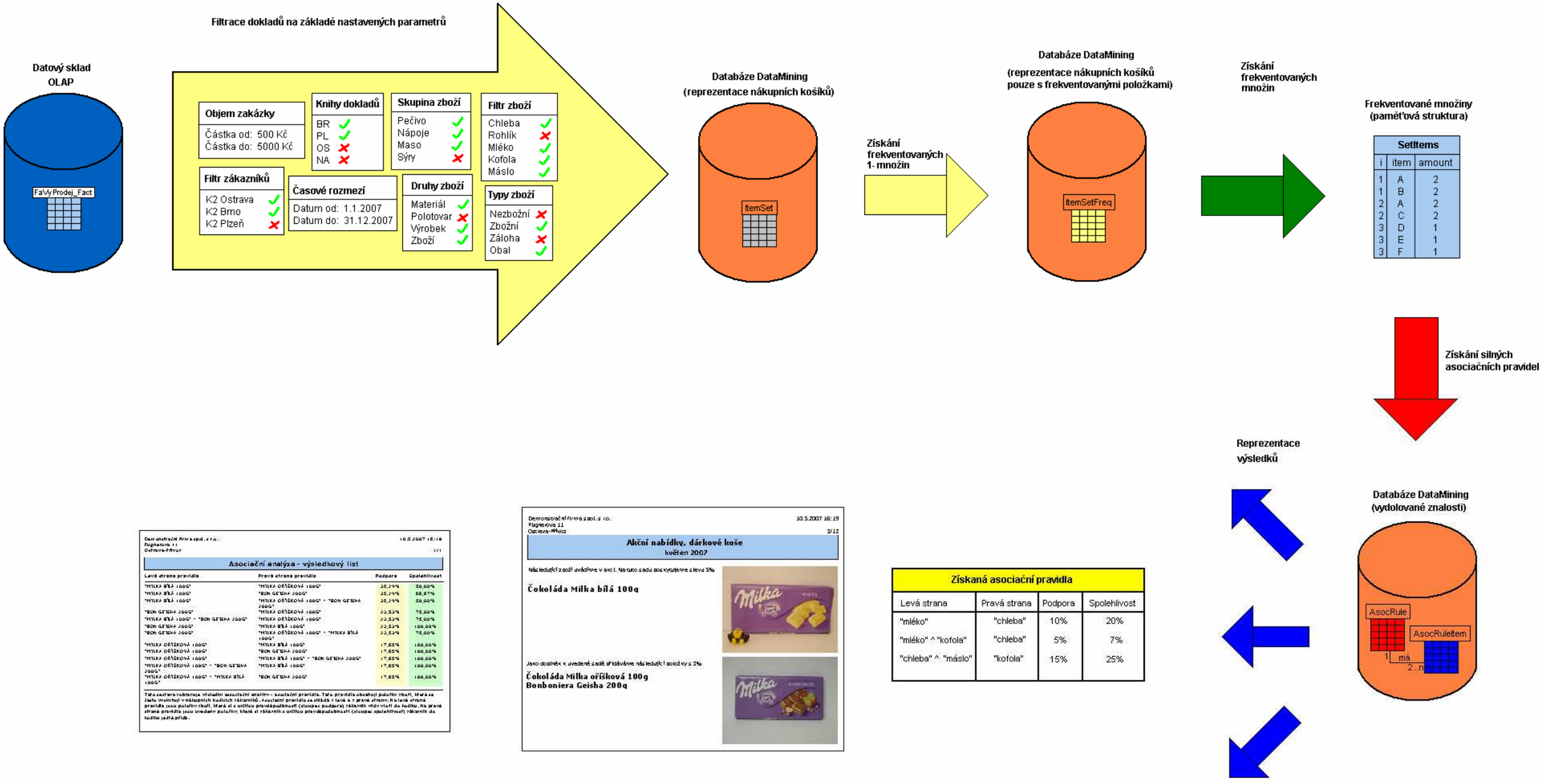
Před počátkem implementace modulu dolování znalostí byl nový modul nejprve rozdělen do tří jednotlivých částí (viz obr. 7). Z hlediska bezpečnosti bylo nutné navrhnout systém práv na přístup k jednotlivým částem modulu a ošetřit možný vstup více uživatelů do určitých částí modulu. V následujících kapitolách detailně popíši implementaci jednotlivých částí modulu a jejich funkci.



Obr. 7 Jednotlivé části modulu

7.1 Návrh modulu

Koncepce řešení uvedená v kapitole 5 je velice zjednodušená. Proto před samotným zahájením implementace bylo nutné navrhnout detailnější schéma, které by důkladně popisovalo jednotlivé kroky dolovacího algoritmu od získání dat, jejich filtraci a zpracování až po reprezentaci výsledků. Navržené schéma je uvedeno na obrázku 8.



Asociační analýza - výsledkový list

Levá strana pravidla	Pravá strana pravidla	Podpora	Spolehlivost
"MILKA BÍLÁ 100G"	"MILKA OŘÍŠKOVÁ 100G"	35,29%	50,00%
"MILKA BÍLÁ 100G"	"BON GEISHA 200G"	35,29%	66,67%
"MILKA BÍLÁ 100G"	"MILKA OŘÍŠKOVÁ 100G" - "BON GEISHA 200G"	35,29%	50,00%
"BON GEISHA 200G"	"MILKA OŘÍŠKOVÁ 100G"	23,53%	75,00%
"MILKA BÍLÁ 100G" - "BON GEISHA 200G"	"MILKA OŘÍŠKOVÁ 100G"	23,53%	75,00%
"BON GEISHA 200G"	"MILKA BÍLÁ 100G"	23,53%	100,00%
"BON GEISHA 200G"	"MILKA OŘÍŠKOVÁ 100G" - "MILKA BÍLÁ 100G"	23,53%	75,00%
"MILKA OŘÍŠKOVÁ 100G"	"MILKA BÍLÁ 100G"	17,65%	100,00%
"MILKA OŘÍŠKOVÁ 100G"	"BON GEISHA 200G"	17,65%	100,00%
"MILKA OŘÍŠKOVÁ 100G"	"MILKA BÍLÁ 100G" - "BON GEISHA 200G"	17,65%	100,00%
"MILKA OŘÍŠKOVÁ 100G" - "BON GEISHA 200G"	"MILKA BÍLÁ 100G"	17,65%	100,00%
"MILKA OŘÍŠKOVÁ 100G" - "MILKA BÍLÁ 100G"	"BON GEISHA 200G"	17,65%	100,00%

Tato sestava zobrazuje výsledky asociační analýzy - asociační pravidla. Tato pravidla obsahují položky zboží, která se často vyskytují v nákupních košících zákazníků. Asociační pravidla se skládá z levé a z pravé strany. Na levé straně pravidla jsou položky zboží, která mí z určitou pravděpodobností (sloupec podpora) zákazník vždy vložit do košíku, na pravé straně pravidla jsou uvedeny položky, které si zákazník s určitou pravděpodobností (sloupec spolehlivost) zákazník do košíku ještě přidá.

Denonsovací firma spol. s r.o. 10.5.2007 16:19
Fügnerova 11
Ostrava-Přívoz 3/12

Akční nabídky, dárkové koše
květen 2007

Následující zboží uvádíme v akci. Na uvedenou sadu poskytujeme slevu 5%

Čokoláda Milka bílá 100g

Jako doplněk k uvedené sadě přidáváme následující položky s 3%

Čokoláda Milka oříšková 100g
Bonboniera Geisha 200g

Získaná asociační pravidla			
Levá strana	Pravá strana	Podpora	Spolehlivost
"mléko"	"chleba"	10%	20%
"mléko" ^ "kofola"	"chleba"	5%	7%
"chleba" ^ "máslo"	"kofola"	15%	25%

Obr. 8 Detailní schéma postupu zpracování dat

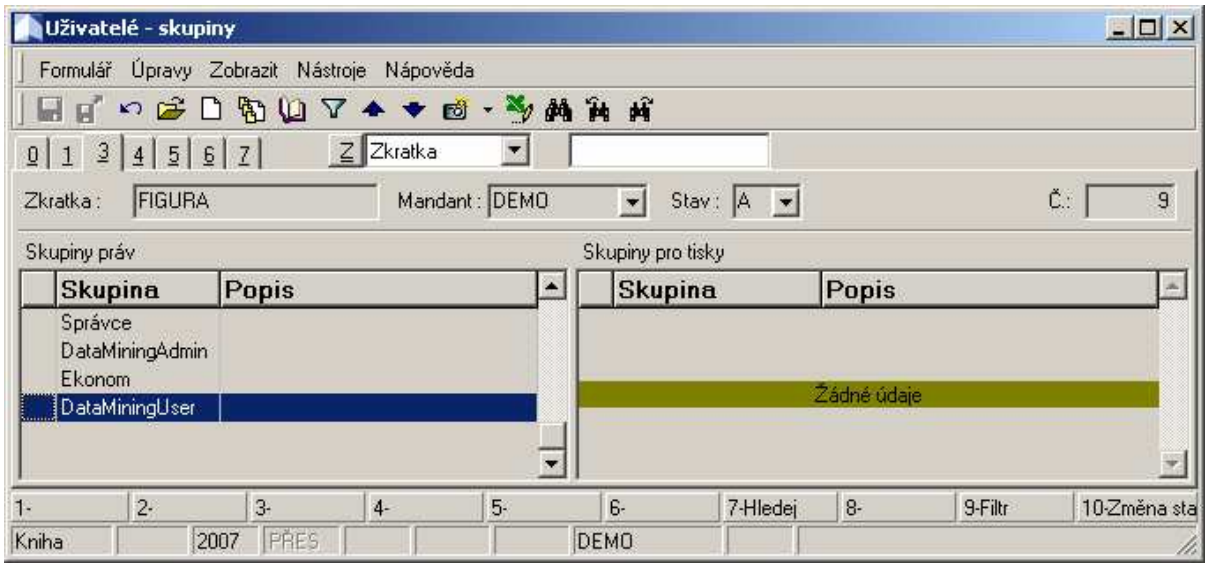
7.2 Přístup k funkcím modulu

7.2.1 Oprávnění

Uživatel informačního systému K2 má přidělenu množinu práv, které ho opravňují k přístupu a změně jeho jednotlivých částí. Obdobně také pro přístup do modulu získávání znalostí byly implementovány dvě skupiny oprávnění:

- **DataMiningAdmin** – oprávnění, které uživateli umožňuje nastavovat parametry dolovacího procesu (přístup k části modulu "nastaven parametrů") a spouštění samotného dolovacího procesu (přístup k části modulu "Dolovací proces") – generování asociačních pravidel.
- **DataMiningUser** – oprávnění, které uživateli umožňuje zobrazovat získané výsledky asociační analýzy – asociační pravidla.

Tyto skupiny oprávnění se nastavují v modulu informačního systému K2 "Správce - uživatelé". Na 3. straně uživatelů je možné přidávat skupiny práv (viz obr. 9).



Obr. 9 Skupiny oprávnění

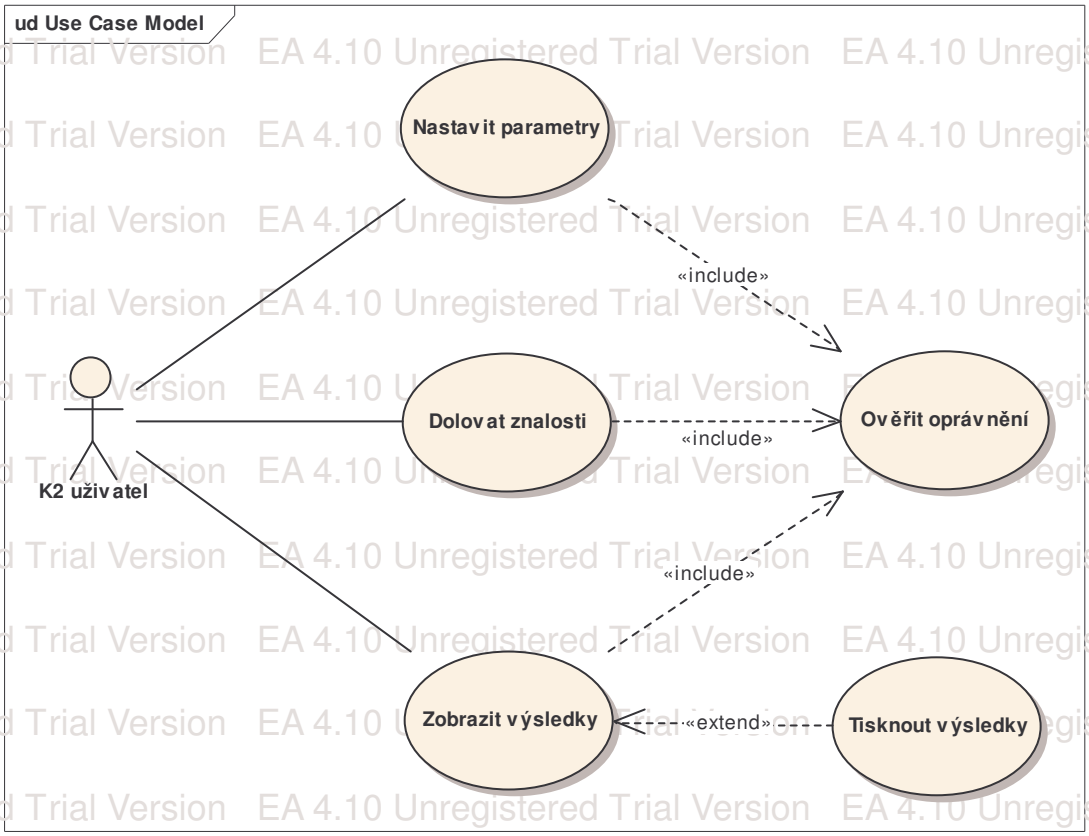
7.2.2 Použití modulu

Modul získávání znalostí tedy umožňuje následující akce:

- **Nastavení parametrů modulu** – v této části uživatel nastavuje parametry, které se vztahují k dolovacímu procesu a systémové parametry modulu (více viz kapitola 7.3.1). Přístup do této části je uživateli povolen pouze v případě, že má nastavenu skupinu oprávnění "DataMiningAdmin".

- **Samotné spuštění dolovacího procesu** – v této části se spustí proces asociační analýzy s nastavenými parametry. Uživatel musí mít nastavenu skupinu oprávnění "DataMiningAdmin".
- **Zobrazení získaných výsledků** – v této části uživatel prohlíží a dále zpracovává získané výsledky dolovacího procesu. Přístup do této části je umožněn pouze uživatelům s nastavenou skupinou oprávnění "DataMiningUser".

Diagram případů použití modulu je uveden na obrázku 10.



Obr. 10 Diagram případů použití modulu

7.2.3 Současné přihlášení více uživatelů

Informační systém K2 je víceuživatelský a jako takový umožňuje souběžnou práci několika uživatelů současně. Koncepce řešení nově vytvářeného modulu pro získávání znalostí počítá s tím, že víceuživatelské přístup k modulu bude umožněn, ovšem všichni uživatelé budou pracovat se stejnou množinou parametrů a získaných výsledků.

Jak již bylo uvedeno výše, modul se skládá ze tří částí: nastavení parametrů, samotný proces dolování a zobrazení získaných výsledků. Do části pro nastavení parametrů může v jednu chvíli současně vstoupit pouze jeden uživatel a to z důvodu, aby byl zachován konzistentní stav parametrů (aby si uživatelé vzájemně neměnili nastavení). Obdobně také do části modulu, která provádí

dolovací algoritmus, smí současně vstoupit pouze jeden uživatel (výsledky se ukládají do databáze, došlo by k přepisu výsledků dříve spuštěného dolovacího procesu výsledky novějšími). Současně vstup více uživatelů do modulu pro prohlížení získaných výsledků není omezen (číst může více uživatelů současně).

Všechny tyto uvedené podmínky musí platit současně s tím, že do části nastavení parametrů smí vstoupit jakýkoliv uživatel pouze v případě, že žádný uživatel není v části dolovacího procesu ani v zobrazení výsledků (obě tyto části jsou na parametrech závislé). Obdobně do části dolovacího procesu smí uživatel vstoupit pouze v případě, že žádný uživatel není v části nastavení parametrů ani v části zobrazení výsledků (v okamžiku nastavování parametrů mohou být tyto parametry nekonzistentní, což by způsobilo havárii části dolování, obdobně zobrazení výsledků). Do části zobrazení výsledků smí uživatel vstoupit pouze v případě, že žádný uživatel není v části dolovacího procesu a ani v nastavení parametrů.

7.2.3.1 Třída cHolder

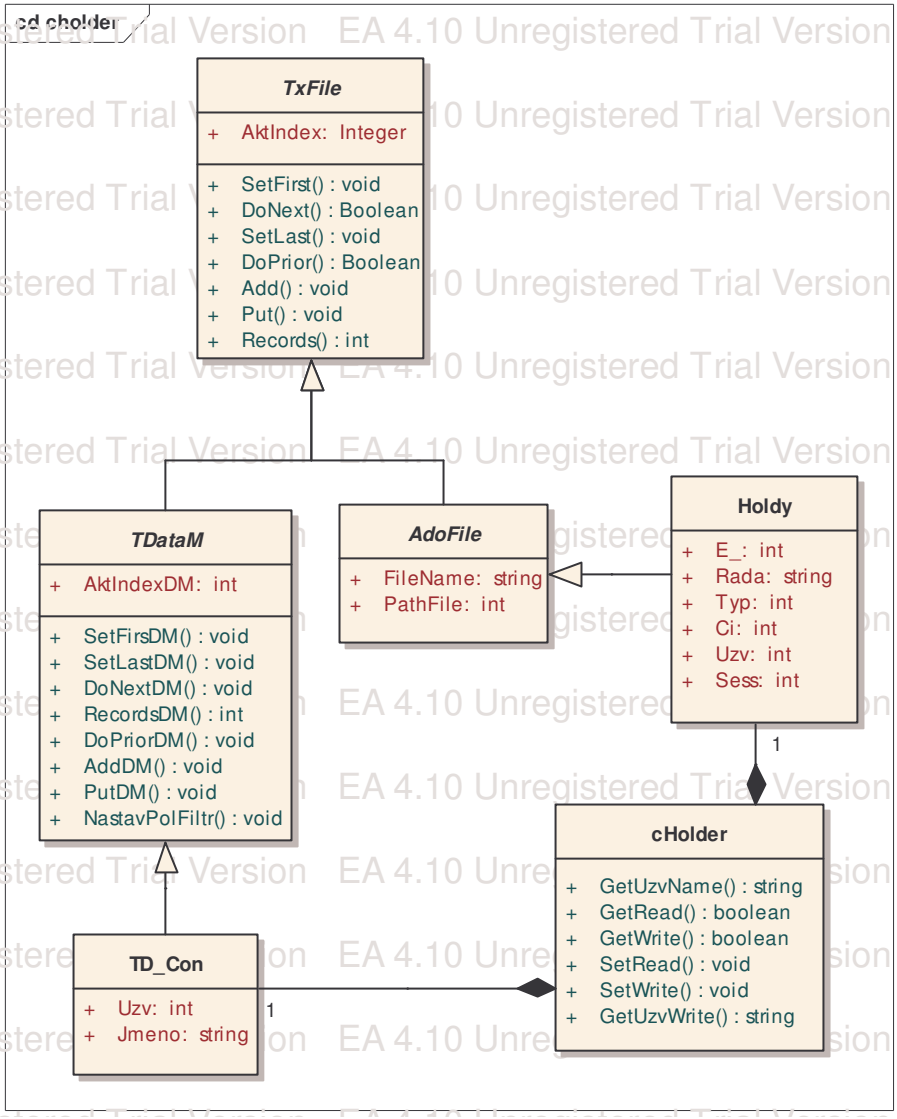
Pro zabezpečení těchto podmínek byla navržena třída cHolder (viz obr. 11), která využívá systémovou tabulku K2 s názvem "Holdy". Tato tabulka je ve skriptu zpřístupněna pomocí třídy "Holdy" (pojmenování tříd pro přístup k tabulkám je shodné jako název tabulky. V této tabulce jsou takzvané držené záznamy, které nemohou ostatní uživatelé měnit. Pro přístup k informacím o uživatelích třída "cHolder" obsahuje také datový modul typu "TD_Con" (více k popisu tříd viz. kapitola 3.2). Tato třída implementuje uzamykací mechanizmus jednotlivých částí modulu.

Pro správně pochopení funkce uzamykání je potřeba definovat dva příznaky: příznak zápisu a příznak čtení. Příznak zápisu je nastaven v okamžiku libovolného zápisu do modulu (zápis parametrů nebo zápis výsledků asociační analýzy). Příznak čtení je nastaven v okamžiku čtení výsledků asociační analýzy.

Postup uzamykání je následující. V okamžiku vstupu uživatele systému do části modulu pro nastavení parametrů se zkontrolují příznaky čtení a změny modulu a v případě, že je vstup umožněn, je do tabulky zapsán příznak o změně modulu. Po ukončení práce s parametry se příznak změny modulu zruší. Obdobný princip je použit pro část modulu pro dolování znalostí.

V části modulu pro zobrazení výsledků se kontroluje pouze příznak změny modulu. V případě, že je vstup umožněn, se inkrementuje (a po ukončení práce dekrementuje) příznak čtení. Tím je ošetřen stav, kdy by do části zobrazení výsledků vstoupil nejprve jeden uživatel, nastavil příznak, poté by vstoupil druhý, který by výsledky opustil a tím by příznak čtení zrušil (a přitom by první uživatel v modulu stále setrvával). Druhý uživatel by následně mohl spustit dolovací algoritmus a tím nevědomky změnit prvnímu uživateli výsledky (příznak čtení by měl totiž vypnutý).

Začlenění třídy cHolder do modulu je zobrazeno na podrobném diagramu tříd viz příloha 3.



Obr. 11 Třída cHolder

7.3 Jednotlivé části modulu

Jak již bylo uvedeno výše, modul se skládá ze tří částí. V této kapitole důkladně popíši tyto jednotlivé části.

7.3.1 Nastavení parametrů

Nejdůležitějším úkolem uživatele dolovacího modulu je správně nastavit parametry dolovací úlohy tak, aby získané výsledky byly v praxi použitelné a aby časová náročnost zpracování dolovacího procesu nepřesáhla rozumnou hranici (například délku víkendu při týdenním zpracování). V neposlední řadě je nutné správně nastavit systémové parametry modulu (tyto parametry se nastaví pouze při nasazení modulu do informačního systému společnosti a dále je uživatel nebude měnit).

7.3.1.1 Parametry asociační analýzy

Základními parametry asociační analýzy jsou hodnoty minimální podpory a minimální spolehlivosti hledaných asociačních pravidel. Pro zajištění ukončení dolovacího algoritmu v případě velni rozsáhlé množiny položek byl přidán parametr na omezení velikosti generované k-množiny (podrobněji viz kapitola 6).

Mezi další parametry patří nastavení omezujících podmínek pro filtraci množiny zakázek (v terminologii K2 se pojem nákupní košík označuje jako zakázka).

Filtrace zakázek:

- **Rozmezí dat** – tato podmínka filtruje zakázky, které nespadají do zadaného rozmezí dat
- **Rozmezí objemu zakázky** – tato podmínka filtruje zakázky, které svou částkou netto nespadají do zadaného rozmezí
- **Filtr knih zakázek** – tato podmínka slouží k nastavení knih zakázek, které se mají v dolovacím procesu uvažovat. Nastavení filtru probíhá výčtem všech evidovaných knih zakázek a jejich povolením nebo zákazem. Kniha zakázky je reprezentací šanonu, do kterého se papírové zakázky vkládají. Takže zakázky, které jsou založeny například v řadě "NA" reprezentují všechny nabídky a obdobně zakázky založené v řadě "BR" mohou reprezentovat zakázky brněnské pobočky firmy.
- **Filtr zákazníků** – tato podmínka slouží k povolení nebo zákazu dolování znalostí ze zakázek vystavených na konkrétní zákazníky. Nastavení filtru probíhá opět výčtem jako v případě filtru knih zakázek.

Další parametry slouží k omezení položek zakázek, čímž se sníží doba potřebná pro zpracování dolovacího algoritmu (prohledává se menší tabulka - irelevantní položky jsou automaticky vynechány).

Filtrace položek:

- **Filtr položek zboží** – obdobně jako v případě filtru knih zakázek je možné také u konkrétních položek zboží explicitně zvolit ty, které nemají vstupovat do dolovacího procesu.
- **Filtr typů zboží** – je možné specifikovat typy zboží, které nemají vstupovat do dolovacího procesu. V informačním systému K2 má každá položka zboží přiřazen svůj typ. Možné typy zboží jsou například "zbožní" nebo "nezbožní".
- **Filtr druhů zboží** – je možné specifikovat druhy zboží, které nemají vstupovat do dolovacího procesu. V informačním systému K2 má každá položka zboží přiřazen svůj druh. Možné druhy zboží jsou například "materiál", "polotovar" nebo "výrobek".

- **Filtr skupin zboží** – je možné specifikovat skupiny zboží, které nemají vstupovat do dolovacího procesu. V informačním systému K2 má každá položka zboží přiřazenu svou skupinu. Možné skupiny zboží jsou například "potraviny", "nářadí" nebo "náhradní díly".
- **Filtr kódů zboží** – je možné specifikovat kódy zboží, které nemají vstupovat do dolovacího procesu. V informačním systému K2 má každá položka zboží přiřazen svůj kód. Možné kódy zboží jsou například "pečivo", "maso" nebo "mléčné výrobky".

7.3.1.2 Systémové parametry modulu

Krom parametrů asociační analýzy je potřeba spravovat systémové parametry modulu. Tyto parametry slouží k nastavení napojení na databázi.

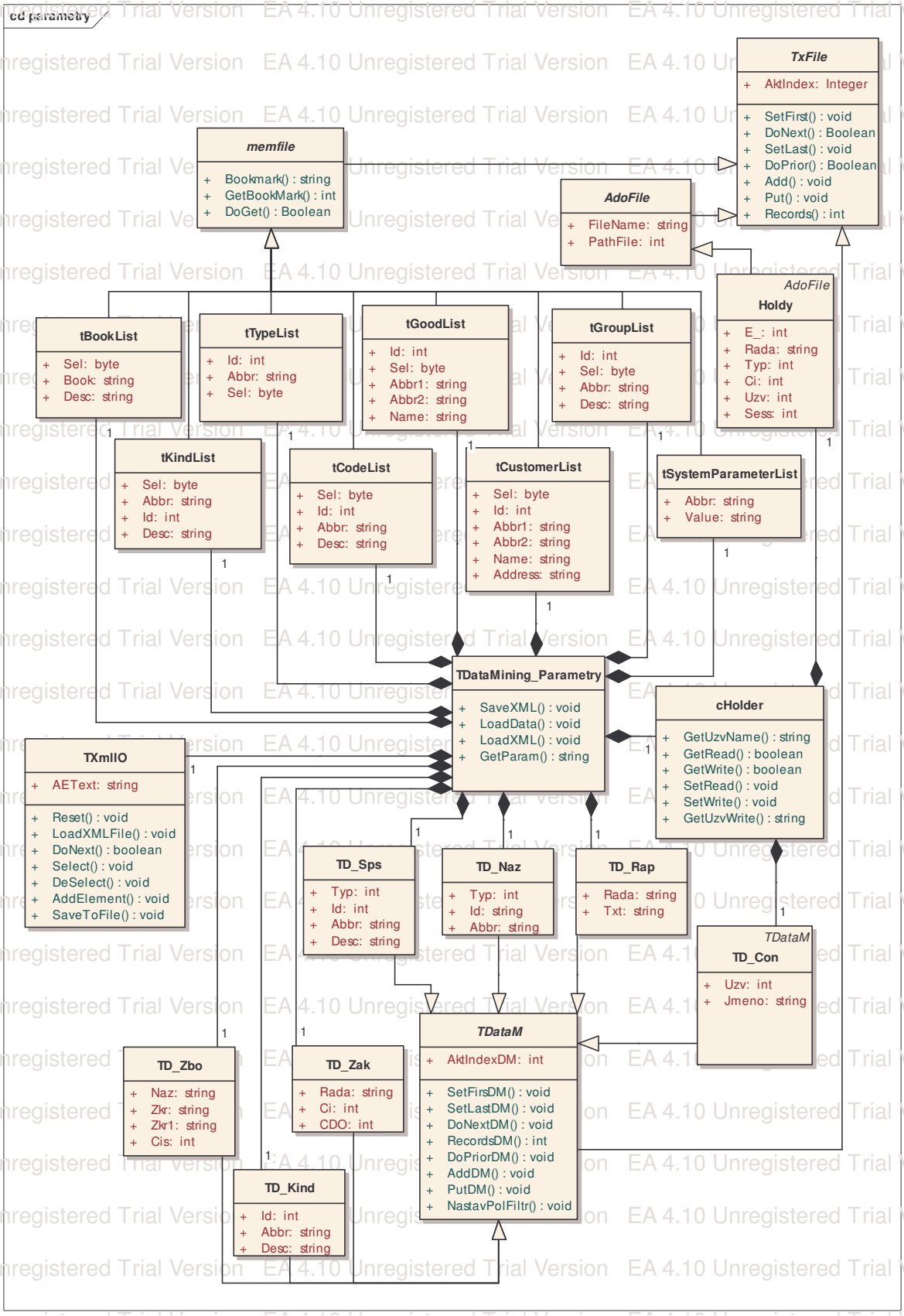
Systémové parametry

- **DBDataMiningName** – přípona databáze modulu dolování dat (implicitně "_DATAMINING")
- **ServerDataminingName** – název serveru s databází modulu (implicitně shodné se serverem K2)
- **DBDataminingUser** – jméno uživatele pro připojení k databází modulu (implicitně "K2")
- **DBWarehouseName** – přípona databáze datového skladu (implicitně "_DM")
- **ServerWarehouseName** – název serveru s datovým skladem (implicitně shodné se serverem K2)
- **DBWarehouseUser** – jméno uživatele pro připojení k databází modulu (implicitně "K2_DM")
- **FactFaVyProdejName** – jméno tabulky datového skladu s daty prodeje (implicitně "FaVyProdej_Fact")
- **ItemSetName** – jméno tabulky modulu s daty nákupních košíků (implicitně "ItemSet")
- **ItemSetFreqName** – jméno tabulky modulu s daty nákupních košíků obsahující pouze frekventované položky (implicitně "ItemSetFreq")
- **LogFileName** – jméno vytvářeného logovacího souboru (implicitně "DataMining.log")
- **AsocRuleName** – jméno tabulky modulu pro ukládání hlaviček asociačních pravidel (implicitně "AsocRule")
- **AsocRuleItemName** – jméno tabulky modulu pro ukládání položek asociačních pravidel (implicitně "AsocRuleItem")

7.3.1.3 Třída parametrů tDataMining_Parametry

Pro správu parametrů dolovacího procesu byla navržena třída TDataMining_Parametry (viz obr. 12). Tato třída obstarává načítání dat informačního systému (např. číselníky, více viz kapitola 7.3.1.1),

načítání a ukládání nastavovaných parametrů a v neposlední řadě také komunikaci s uživatelem. Umístění třídy parametrů v modulu je podrobně znázorněna na diagramu tříd viz příloha 3.



Obr. 12 Třída TDataMining_Parametry

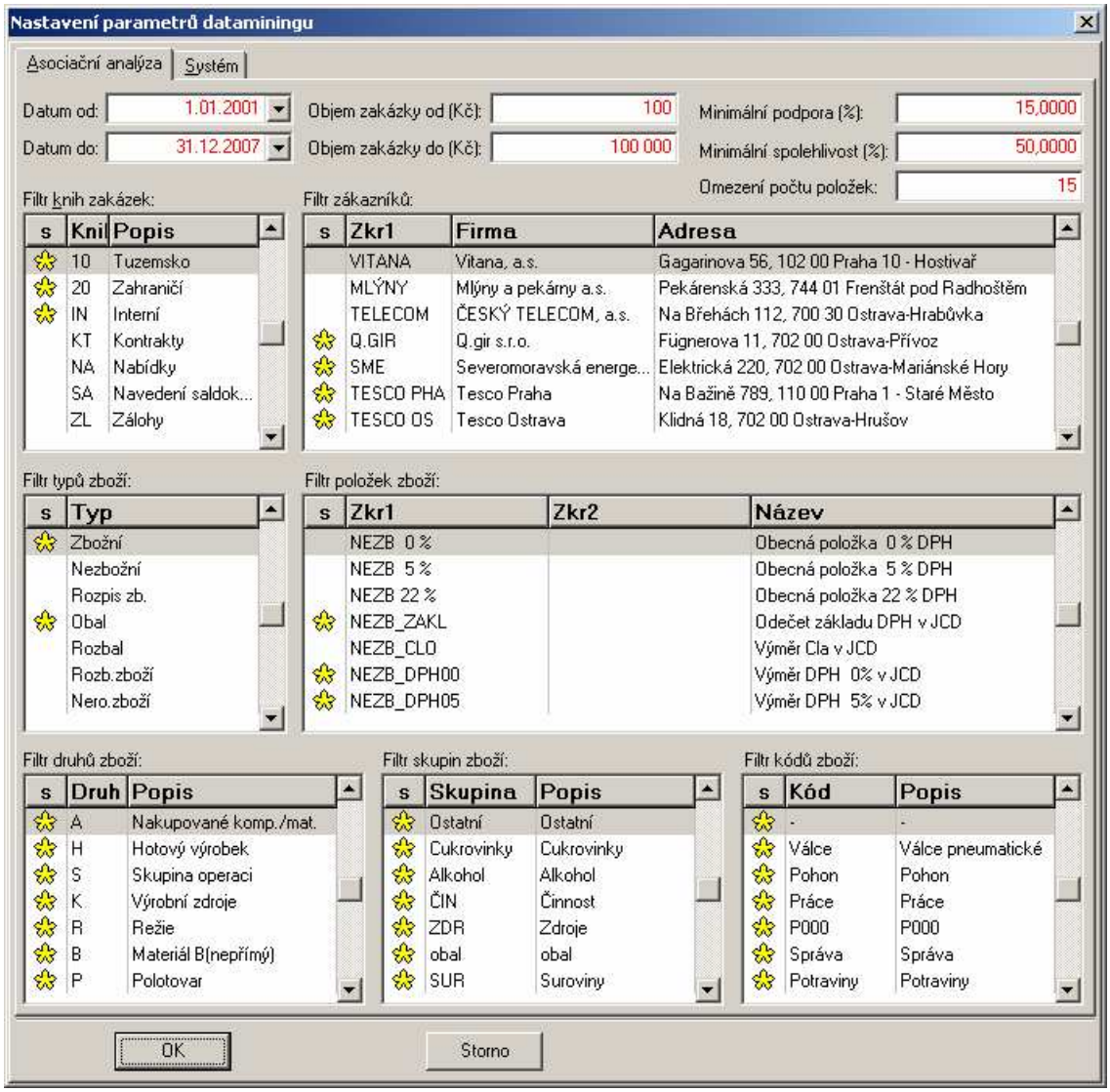
Parametry uvedené v kapitole 7.3.1.1 jsou v třídě parametrů načítány do jejich atributů. Tyto atributy jsou v případě číselníků (např. nastavení knih dokladů) typu memfile (více viz kapitola 3.2). Tento typ je třída obsahují kolekci záznamů a metody pro pohyb na těchto záznamech a jejich změnu. Hodnoty číselníků (data K2) jsou v třídě zpřístupněna pomocí následujících tříd:

- "TD_Sps" - společný číselník pro skupiny a kódy zboží
- "TD_Rap" – knihy prodeje
- "TD_Naz" – číselník typů zboží
- "TD_Kind" – číselník druhů zboží
- "TD_Zbo" – číselník položek zboží
- "TD_Zak" – číselník jednotlivých zákazníků

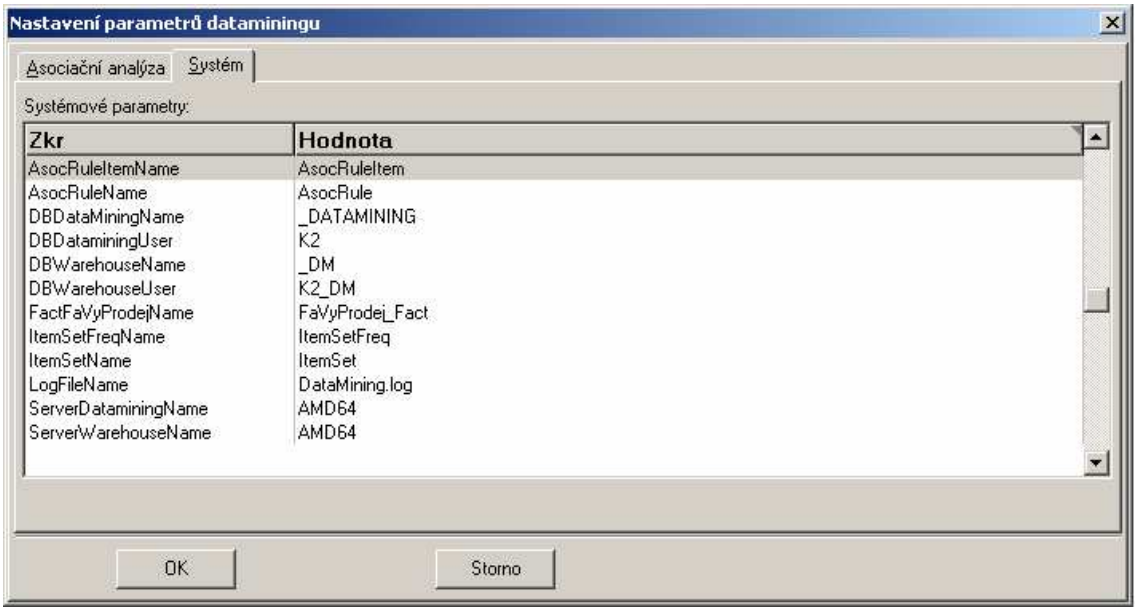
Parametry jednoduchých datových typů jsou ukládány přímo do formulářových komponent uživatelského rozhraní. Seznam struktur (memfile) číselníků a jejich popis je uveden v příloze 1. Pro načítání a ukládání dat z XML souboru třída "TDataMining_Parametry" obsahuje třídu typu "TXmlIO" a pro zjištění a nastavení stavu uzamknutí modulu (více viz kapitola 7.2.3) obsahuje třídu "cHolder".

Návrh uživatelského rozhraní

Třída parametrů dolovacího modulu obsahuje uživatelské rozhraní ve formě formuláře. Tento formulář obsahuje dvě záložky. První slouží k nastavení parametrů asociační analýzy (viz obr. 13) a druhá k nastavení systémových parametrů (viz obr. 14).



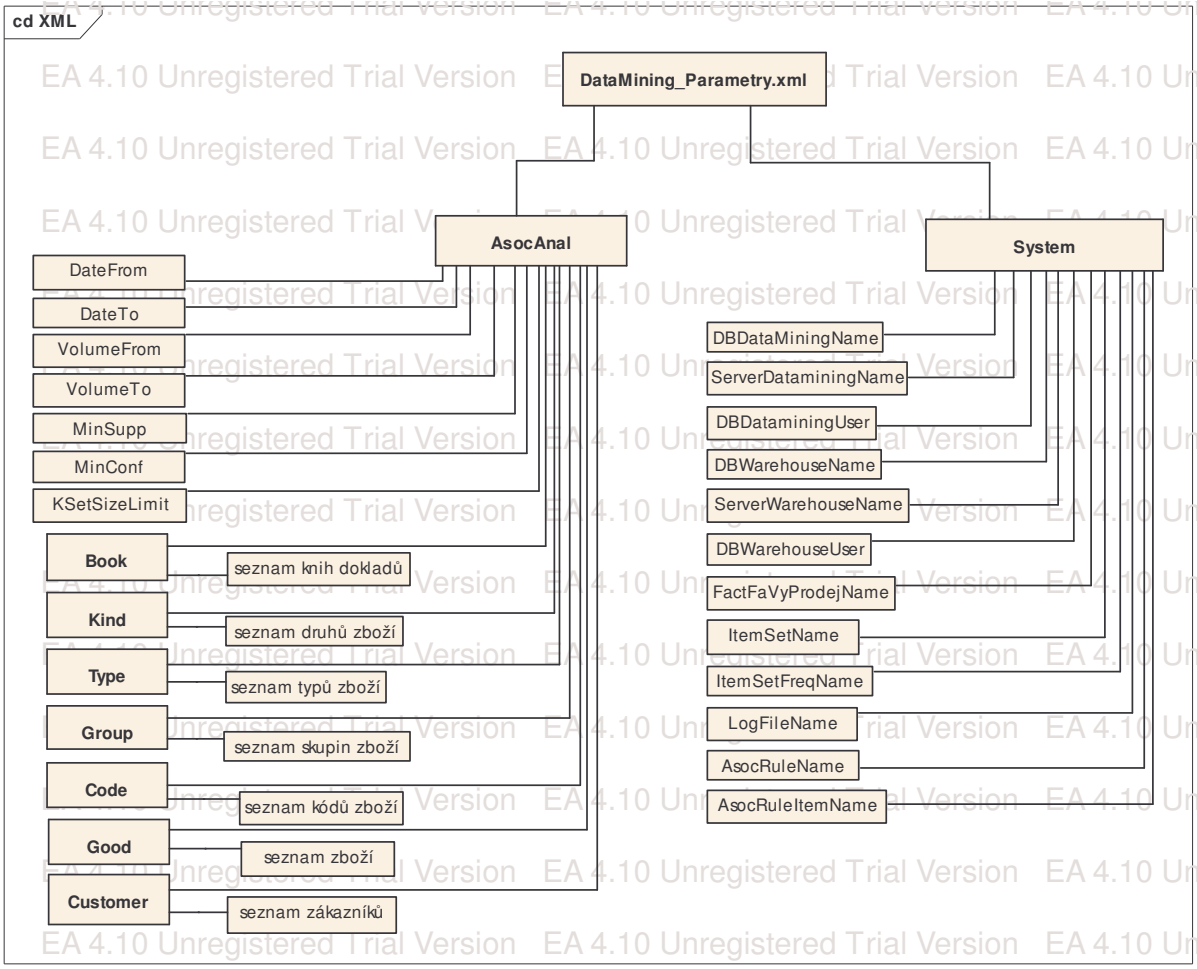
Obr. 13 Parametry asociační analýzy



Obr. 14 Systémové parametry modulu

Uložení parametrů

Nastavené parametry je potřeba uložit, pro uložení parametrů jsem zvolil dokument XML. Struktura souboru je znázorněna na obrázku 15. XML dokument byl zvolen z důvodu snazší přenositelnosti nastavení parametrů mezi různými zákazníky.



Obr. 15 XML struktura souboru parametrů

7.3.2 Dolovací proces

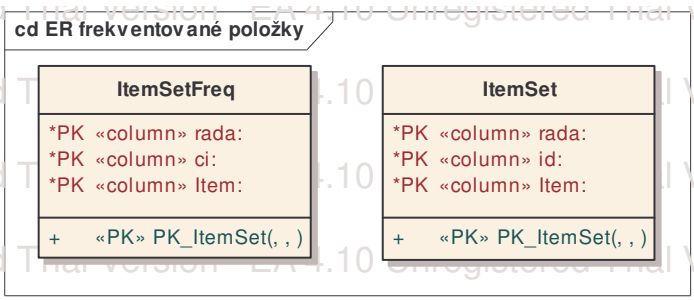
Samotný proces dolování je časové velice náročná operace, a proto je zcela uživatelsky neinteraktivní. Uživatel má možnost pouze sledovat průběh procesu a to buď přímo (viz obr. 23) nebo zpětně z logovacího souboru (viz obr. 22).

7.3.2.1 Kopie dat do databáze modulu dolování znalostí

V prvním kroku dolovacího procesu se načítají data z datového skladu K2. Tato data se filtrují podle nastavených parametrů (viz kapitola 7.3.1). Parametry se vyhodnocují všechny postupně na každou položku (položka je zkopírována pouze v případě, že vyhovuje průniku všech pravidel). Ze zdrojové tabulky se vybírají pouze relevantní informace (sloupce) a kopírují se do tabulky modulu dolování

znalostí. Pro typ úlohy dolování asociačních pravidel stačí znát v množině položek nákupních košíků pouze identifikaci nákupního košíku, kde daná položka náleží a identifikaci této položky. Jednoznačná identifikace nákupního košíku (zakázky) je v informačním systému K2 zabezpečena pomocí knihy dokladu zakázky a čísla zakázky. Identifikací položky zakázky je číslo zboží. Tyto tři hodnoty tedy tvoří záznam jedné položky nákupního košíku (tvoří současně unikátní index na tabulce). V informačním systému K2 je umožněno vložení stejných položek zboží na jednu zakázku. Proto jsou stejné položky zboží při kopii do databáze modulu seskupeny. Data jsou kopírována do tabulky "ItemSet" (implicitní název, konkrétní název závisí na nastavení parametru dolovací úlohy). Struktura tabulky položek nákupních košíků je uvedena na obrázku 16.

Pozn.: Pro tvorbu ER diagramů byl použit program Enterprise Architect, který pro primární klíče používá notaci, která se podobá notaci metod tříd.



Obr. 16 Struktura tabulek položek nákupních košíků

7.3.2.2 Získání frekventovaných položek

V tomto kroku jsou již položky nákupních košíků (položky zakázek) uloženy v databázi modulu dolování znalostí. Pro urychlení prohledávání této databáze byla zvolena jedna dodatečná filtrovací podmínka – omezení na frekventované položky. Pro každou položku je zjištěn procentuální podíl výskytů této položky ve všech zakázkách a pokud je tato hodnota větší nebo rovna minimální podpoře, tak je tato položka zkopírována do vedlejší tabulky. Struktura této tabulky je shodná s tabulkou položek nákupních košíků. Tímto se výrazně zmenší počet záznamů tabulky frekventovaných položek. Ohodnocení množiny položek na počet výskytů ve všech nákupních košících zabezpečuje sql dotaz nad tabulkou položek nákupních košíků (viz obr. 17).

```
SELECT COUNT(*) AS c, rada, ci
  FROM ItemSetFreq
  WHERE  (item = 24) OR
        (item = 42) OR
        (item = 53)
  GROUP BY rada, ci
  HAVING (COUNT(*) > 2)
```

Obr. 17 Ukázka sql dotazu pro ohodnocení počtu výskytů položek

Je založen na agregační funkci "count" a s výhodou využívá klauzuli "having". V klauzuli "where" jsou uvedeny položky, u kterých se zjišťuje počet výskytů (v uvedeném případě se zjišťuje počet výskytů v nákupních košících, které obsahují všechny tři položky současně – 24, 42 a 53). Podmínka na počet v klauzuli "having" odpovídá počtu položek (počet výskytů musí být minimálně stejný, jako počet položek). Vzhledem k tomu, že identifikátor položky "item" (id zboží systému K2) je v rámci skupiny položek identifikovaných sloupci "rada" (kniha dokladu) a "ci" (číslo dokladu) unikátní, lze tento sql dotaz využít ke zjištění počtu výskytů položek ve všech nákupních košících.

7.3.2.3 Získání frekventovaných množin

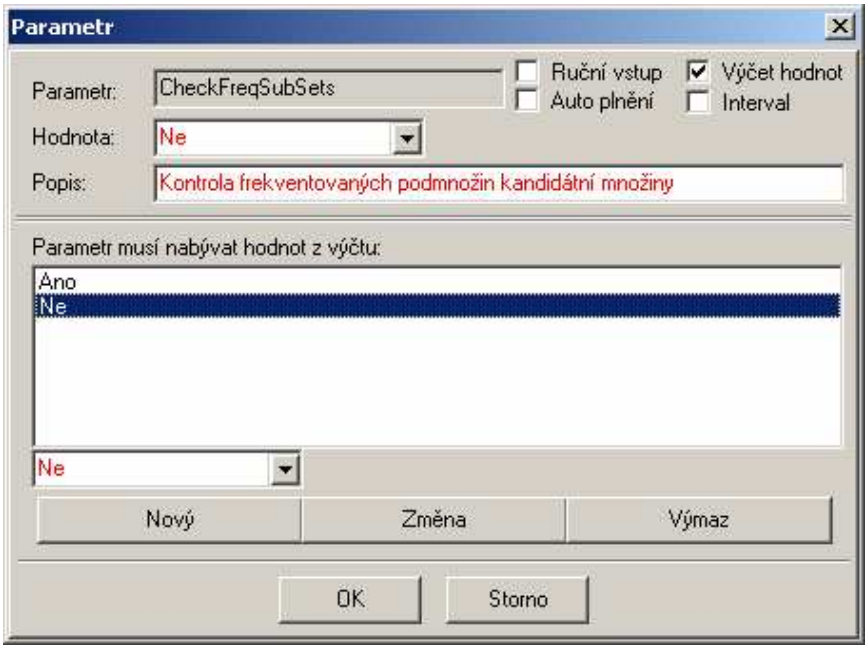
V tomto okamžiku jsou v tabulce "ItemSetFreq" nákupní košíky obsahující pouze frekventované položky a z nich se jediným sql dotazem získají frekventované 1-množiny. Získání dalších frekventovaných k-množin je založeno na algoritmu apriori (viz kapitola 6). Cyklicky se opakuje slučovací a vylučovací krok na množině frekventovaných k-množin a to tak dlouho, dokud je počet frekventovaných množin neprázdný nebo dokud velikost frekventované k-množiny nedosáhla limitu nastaveného parametrem "Omezení počtu položek" (viz kapitola 7.3.1). Detailní popis slučovacího a vylučovacího kroku je uveden v kapitole 6.

Skriptovací jazyk informačního systému K2 neobsahuje datové struktury pro efektivní práci s množinami. Proto byla implementace založena na strukturách typu "memfile". Tyto struktury obsahují kolekci záznamů (deklarovaného typu) a metody pro přidání a odstranění prvku a pro pohyb po těchto záznamech. Množina byla tedy definována jako skupina záznamů kolekce, kde každá položka množiny obsahovala stejný identifikátor (kterým byla množina identifikována).

Slučovací krok byl implementován jako dvojitý cyklus, v jehož vnější části se načítalo k-1 položek postupně všech množin a v jehož vnitřní části se množina položek z vnější části spojovala se všemi ostatními množinami stejné kolekce. Ke spojení množin došlo, když jejich prvních k-2 položek bylo shodných.

Nově vytvořená k-množina může být do frekventovaných množin přidána pouze v případě, že je frekventovaná. Dle algoritmu apriori je množina položek frekventovaná, pokud jsou její podmnožiny také frekventované. Implementace této vlastnosti vyžaduje ukládání všech dříve frekventovaných množin (každé frekventované množině je vypočten hash). Pro kontrolu frekventovanosti podmnožin kandidátní množiny postačuje vygenerování všech jejich k-1 podmnožin. U těchto podmnožin se testuje, jestli jsou frekventované (kontroluje se, jestli jsou všechny obsaženy v dříve nalezených frekventovaných množinách) pomocí hashovací funkce. Výpočet hodnoty hash pro množinu položek probíhá tak, že se ze všech identifikátorů položek této množiny složí textový řetězec, který se pomocí funkce MD5 převede na hash. Hash hodnoty všech frekventovaných množin tvoří index na memfile, ve kterém jsou uloženy, takže zjištění, jestli je nějaké množina obsažena ve frekventovaných množinách je velice rychlá. Experimentálně bylo zjištěno, že tento způsob testování kandidátní množiny položek je pro relativně malé databáze

pomalejší (z důvodu potřeby generování všech k-1 podmnožin kandidátní množiny) než přímé otestování počtu výskytů množiny položek v nákupních košících pomocí sql dotazu viz obr. 17. Proto byl do procesu dolování asociačních pravidel přidán parametr "CheckFreqSubSets", pomocí kterého se přepíná, jestli má být před samotným otestováním množiny na počet výskytů v nákupních košících množina otestována na frekventované podmnožiny (příklad vložení parametru na 2. stranu zařazení skriptu je uveden na obrázku 18). Tento parametr je možné zapnout v případě velké tabulky položek nákupních košíků, kdy by vyhodnocení ohodnocovacího sql dotazu (viz obr. 17) trvalo déle, než testování podmnožin na frekventované podmnožiny.



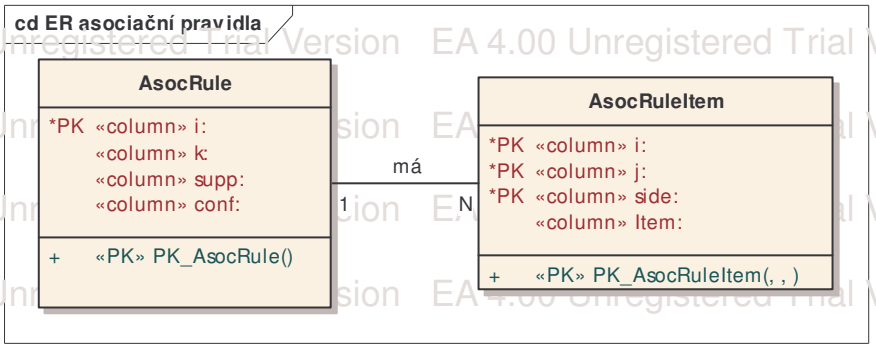
Obr. 18 Parametr na kontrolu dříve nefrekventovaných podmnožin

7.3.2.4 Získání silných asociačních pravidel

V tomto okamžiku jsou získány frekventované množiny. Následuje generováni silných asociačních pravidel. Generování silných asociačních pravidel se skládá z vytvoření všech neprázdných podmnožin frekventovaných množin položek, u kterých se zjistí jejich podpora (všechny jsou frekventované, protože vznikly z frekventovaných množin). Poté se ze získaných podmnožin vytvoří asociační pravidla tak, že na levě straně asociačního pravidla jsou položky podmnožiny a na pravé straně jsou položky doplňku uvedené podmnožiny do frekventované množiny. U takto získaných asociačních pravidel se vypočte jejich spolehlivost pomocí vzorce (3).

Získaná asociační pravidla jsou uložena do databáze ve formě hlavičky a jejich položek (viz obr. 19). V hlavičce asociačního pravidla jsou uvedeny informace o podpoře (sloupec "supp"), spolehlivosti (sloupec "conf") a počtu položek (zjistí se počet položek levé a pravé strany a větší hodnota se použije - má význam pro následné filtrování asociačních pravidel, sloupec "k"). Položky asociačního pravidla jsou uloženy s příznakem "side", který říká, na kterou stranu pravidla položka

náleží. Je důležité upozornit, že minimální počet položek hlavičky asociačního pravidla je dvě (jedna na levé a jedna na pravé straně).



Obr. 19 ER diagram tabulek získaných asociačních pravidel

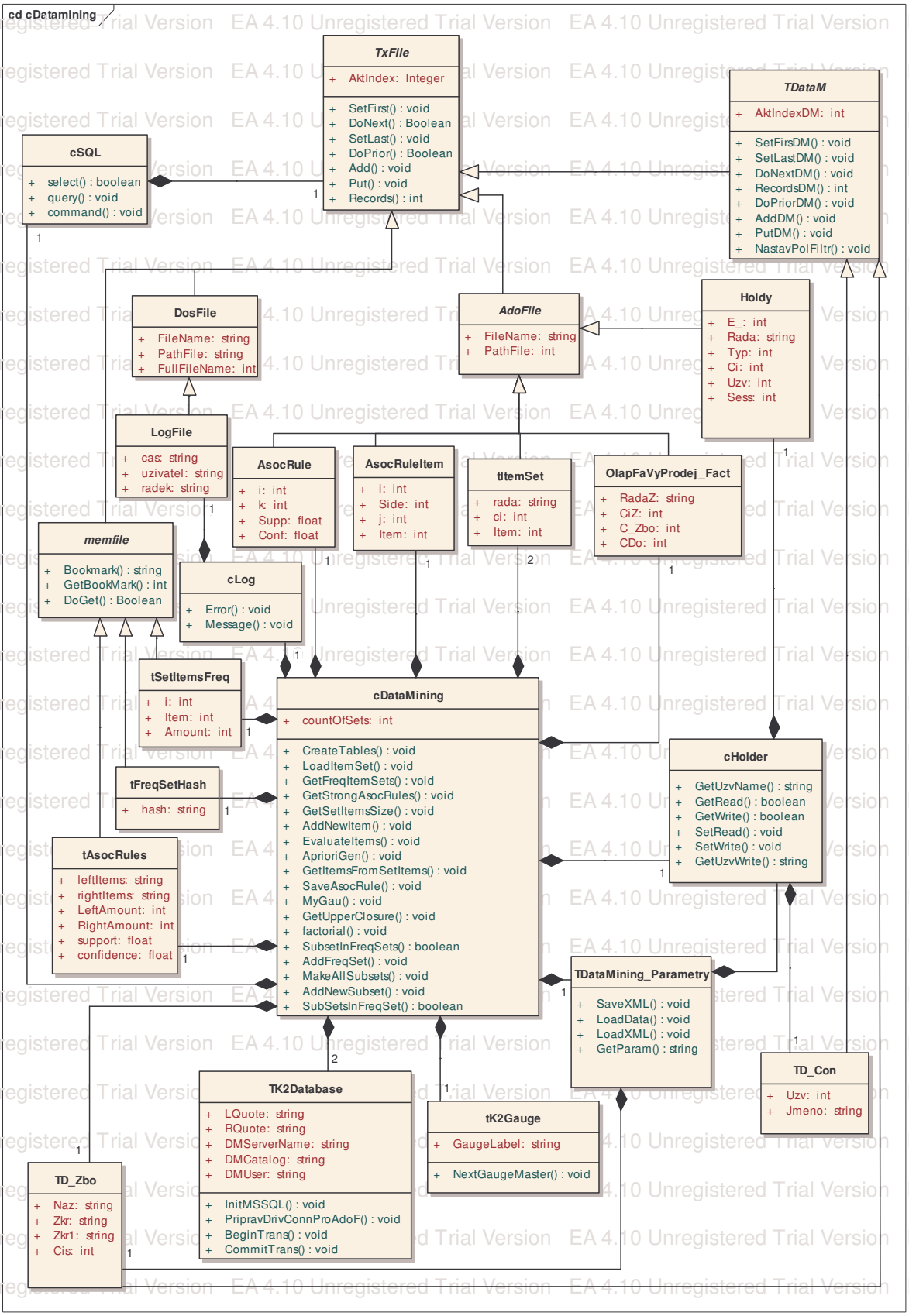
7.3.2.5 Třída dolovacího procesu cDataMining

Třída dolovacího procesu "cDatamining" (viz obr. 20) obsahuje metody pro ovládání procesu dolování znalostí, pomocné metody pro práci s množinami položek, metody pro ukládání získaných výsledků a v neposlední řadě metody pro zobrazování průběhu procesu uživateli.

Třída "cDatamining" obsahuje třídy, jejichž funkce bude nyní vysvětlena:

- "TDataMining_Parametry" - slouží pro načítání nastavených parametrů dolovací úlohy a systémových parametrů pro napojení procesu dolování na databázi (více viz kapitola 7.3.1.3)
- "TD_Zbo" - slouží pro zjištění parametrů položky zboží pro účely filtrování položek nákupních košíků v okamžiku kopírování položek z datového skladu do tabulky nákupních košíků
- "cHolder" – slouží pro zjištění a nastavení stavu uzamknutí modulu
- "cLog" – slouží pro zaznamenání průběhu zpracování dolovacího procesu do logovacího souboru
- "cSQL" – slouží pro efektivní práci s SQL dotazy
- "tK2Gauge" – slouží pro zobrazení průběhu dolovacího procesu
- "TK2Database" – slouží pro napojení modulu na datový sklad a na databázi modulu získávání znalostí
- "tAsocRules" – jedná se o memfile pro uchovávání generovaných asociačních pravidel
- "tFreqSetHash" – slouží pro uchovávání hash hodnot získaných frekventovaných množin
- "tSetItemsFreq" - slouží pro uchovávání získaných frekventovaných množin
- "AsocRule" – třída pro práci s tabulkou hlaviček asociačních pravidel
- "AsocRuleItem" - třída pro práci s tabulkou asociačních pravidel
- "tItemSet" – třída pro práci s tabulkou položek nákupních košíků a s tabulkou frekventovaných položek nákupních košíků

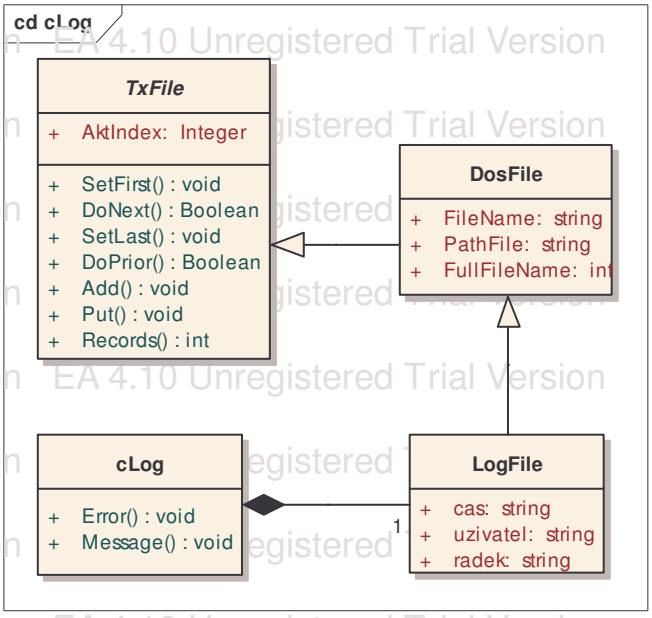
- "OlapFaVyProdej_Fact" – třída pro práci s faktovou tabulkou s daty nákupních košíků v datovém skladu



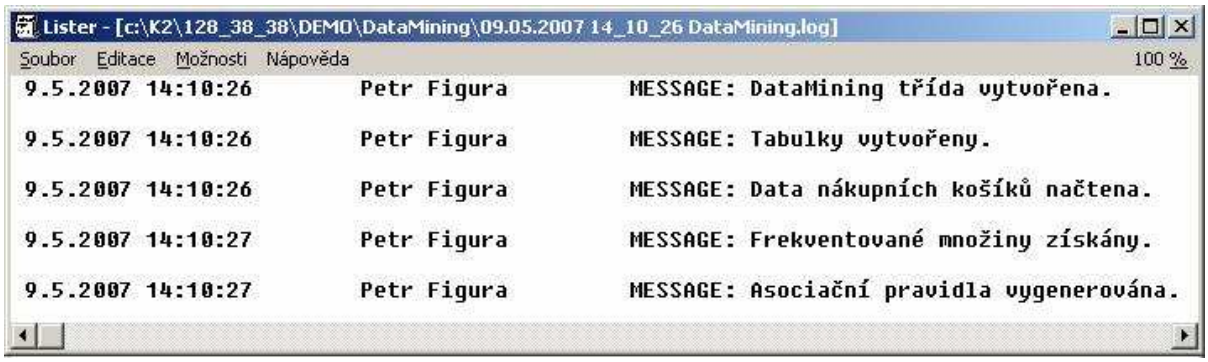
Obr. 20 Třída cDataMining

Logování procesu dolování znalostí

Pro zpětné zobrazení průběhu dolovacího procesu byla vytvořena třída "cLog" (viz obr. 21). Tato třída obsahuje metody pro vytvoření logovacího souboru (s názvem dle parametru viz kapitola 7.3.1), pro zápis běžného a chybového hlášení. Ukázka obsahu logovacího souboru viz obr. 22.



Obr. 21 Třída cLog



Obr. 22 Logovací soubor

7.3.2.6 Uživatelské rozhraní – zobrazení průběhu procesu dolování

Proces dolování znalostí je zcela neinteraktivní. Uživatel má možnost pouze sledovat průběh dolovacího procesu. Po spuštění dolovacího procesu se uživateli zobrazí formulář ukazatele průběhu (viz obr. 23), který se skládá z následujících kroků:

1. Vytváření tabulek
2. Plnění tabulek
3. Získání frekventovaných množin – příprava
4. Získání frekventovaných množin

5. Získání silných asociačních pravidel - generování podmnožin
6. Získání silných asociačních pravidel

Obr. 23 Uživatelské rozhraní – zobrazení průběhu procesu

Odhad doby trvání jednotlivých kroků byl stanoven na základě výpočtu kombinačního čísla viz vzorec (4). Jedná se o variaci k-té třídy z n prvků bez opakování. Odhad na základě výpočtu určuje horní možnou hranici počtu kroků, v praxi však průběžně dochází k vylučování kandidátních množin, takže počet kandidátních množin v další iteraci algoritmu apriori je menší, než odhadovaná hodnota. Odhad je však potřeba provést, aby byl uživatel při sledování průběhu procesu informován o možné délce procesu a především o "živosti" procesu.

$$V_k(n) = \frac{n!}{(n-k)!} \qquad (4)$$

7.3.3 Zobrazení výsledků dolování

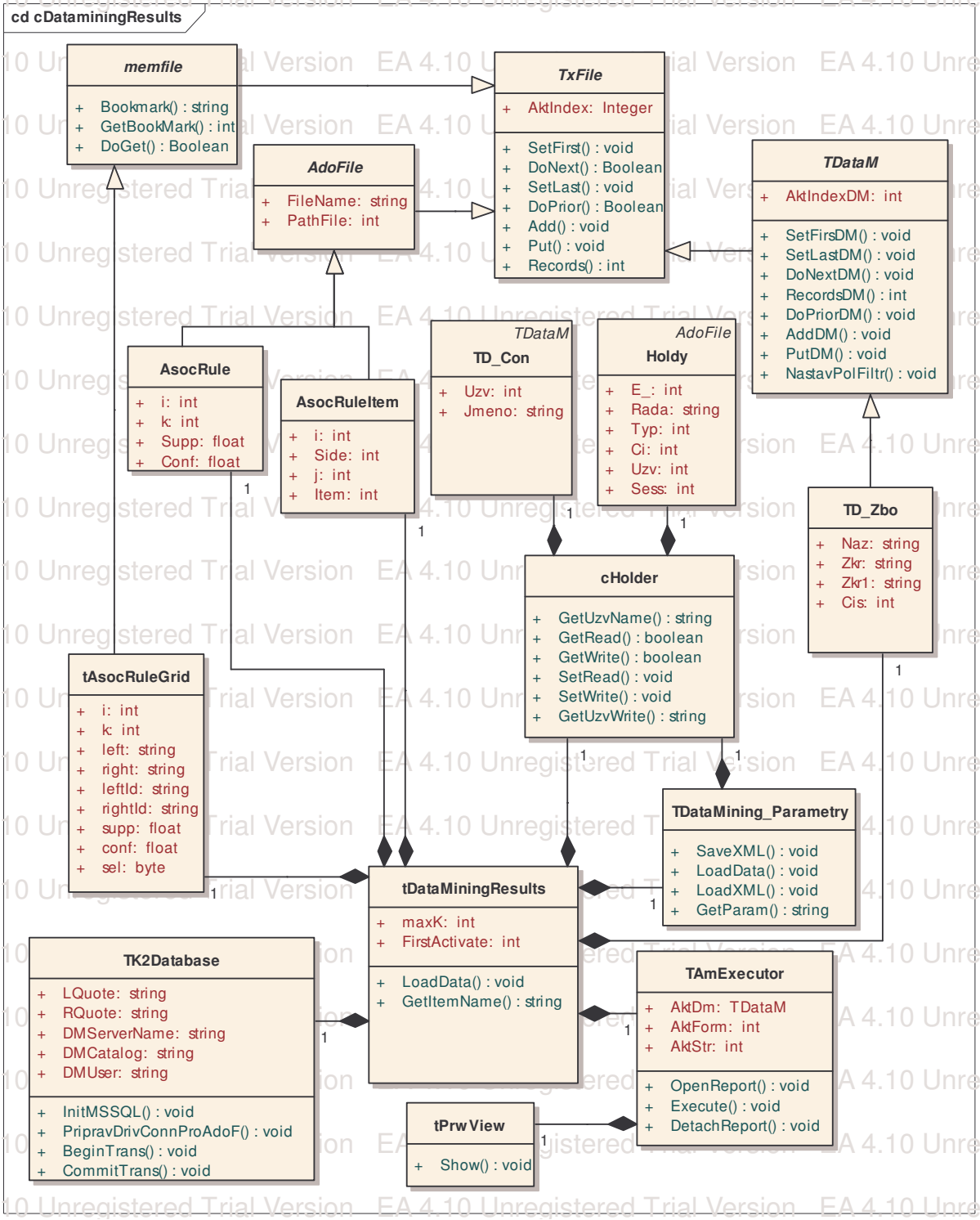
Asociační pravidla - výsledky asociační analýzy jsou uloženy v databázi modulu získávání znalostí. K zobrazení výsledků slouží část modulu "Výsledky asociační analýzy".

7.3.3.1 Třída tDataMiningResult

Pro práci s výsledky asociační analýzy byl navržena třída (viz obr. 24). Tato třída umožňuje načtení hlaviček a položek asociačních pravidel z databáze a následně jejich převedení do textové podoby srozumitelné pro uživatele. Položky na obou stranách pravidla reprezentují zboží evidované v informačním systému K2. Třída tDataminingResults obsahuje následující třídy:

- "TD_Zbo" – slouží pro změnu zobrazení asociačního pravidla – načtení atributů zboží při přepínání mezi identifikátorem, názvem a zkratkami položky zboží (viz kapitola 7.3.3.2)
- "AsocRule" – třída pro práci s tabulkou hlaviček asociačních pravidel
- "AsocRuleItem" – třída pro práci s tabulkou položek asociačních pravidel
- "cHolder" – slouží pro zjištění a nastavení stavu uzamknutí modulu

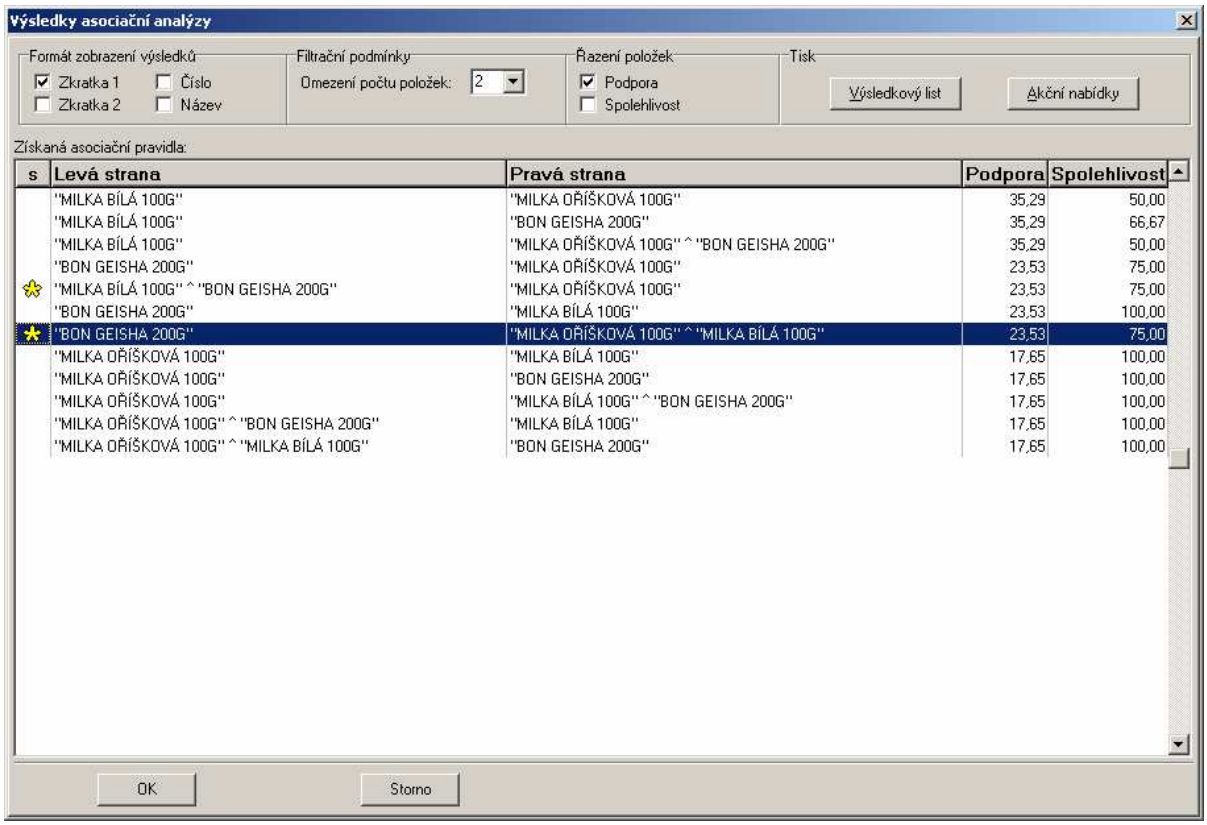
- "TDatamining_Parametry"- slouží pro načítání nastavených parametrů dolovací úlohy a systémových parametrů pro napojení procesu dolování na databázi (více viz kapitola 7.3.1.3)
- "tAsocRuleGrid" – memfile asociačních pravidel pro zobrazení v gridu
- "TK2DatabaseTAmExecutor" – slouží pro vytváření tiskových sestav



Obr. 24 Třída tDataMinigResults

7.3.3.2 Uživatelské rozhraní

Formulář uživatelského rozhraní (viz obr. 25) zobrazuje získaná asociační pravidla jako řádky tabulky. Každý záznam obsahuje levou a pravou stranu asociačního pravidla, jeho podporu a spolehlivost.



Obr. 25 Uživatelské rozhraní modulu

Zobrazení formátu asociačních pravidel je možné měnit zaškrtnutím příslušné volby. Uživatel má na výběr:

- Zkratka 1 - zkratka 1 položky zboží informačního systému K2
- Zkratka 2 - zkratka 2 položky zboží informačního systému K2
- Číslo - interní číslo položky zboží informačního systému K2
- Název - název položky zboží informačního systému K2

Při získání velkého počtu asociačních pravidel je možné dodatečně omezit jejich výpis tím, že si uživatel zvolí maximální počet položek asociačního pravidla. Získaná asociační pravidla je možné řadit buď podle jejich minimální podpory nebo minimální spolehlivosti.

7.3.3.3 Tiskové sestavy

Třída výsledků asociační analýzy umožňuje ze získaných asociačních pravidel vytvářet tiskové sestavy. Data těchto tiskových sestav se načítají z formuláře zobrazení výsledků. Výběr asociačních pravidel, které chce uživatel tisknout se provádí pomocí označení pravidel mezerníkem (zobrazí se

symbol hvězdičky) nebo pomocí kláves "+" a "-" na numerické klávesnici (což ovlivňuje označení všech pravidel najednou).

Výsledkový list

Tato tisková sestava zobrazuje přímé výsledky asociační analýzy - asociační pravidla s hodnotami jejich podpory a spolehlivosti (viz obr. 26). Formát zobrazení asociačních pravidel v tiskové sestavě odpovídá zvolenému způsobu na formuláři.

Demonstrační firma spol. s r.o.
Fügnerova 11
Ostrava-Přívoz

11.5.2007 17:44
1/1

Asociační analýza - výsledkový list

Levá strana pravidla	Pravá strana pravidla	Podpora	Spolehlivost
"MILKA BÍLÁ 100G"	"MILKA OŘÍŠKOVÁ 100G"	35,29%	50,00%
"MILKA BÍLÁ 100G"	"BON GEISHA 200G"	35,29%	66,67%
"MILKA BÍLÁ 100G"	"MILKA OŘÍŠKOVÁ 100G" ^ "BON GEISHA 200G"	35,29%	50,00%
"BON GEISHA 200G"	"MILKA OŘÍŠKOVÁ 100G"	23,53%	75,00%
"MILKA BÍLÁ 100G" ^ "BON GEISHA 200G"	"MILKA OŘÍŠKOVÁ 100G"	23,53%	75,00%
"BON GEISHA 200G"	"MILKA BÍLÁ 100G"	23,53%	100,00%
"BON GEISHA 200G"	"MILKA OŘÍŠKOVÁ 100G" ^ "MILKA BÍLÁ 100G"	23,53%	75,00%
"MILKA OŘÍŠKOVÁ 100G"	"MILKA BÍLÁ 100G"	17,65%	100,00%
"MILKA OŘÍŠKOVÁ 100G"	"BON GEISHA 200G"	17,65%	100,00%
"MILKA OŘÍŠKOVÁ 100G"	"MILKA BÍLÁ 100G" ^ "BON GEISHA 200G"	17,65%	100,00%
"MILKA OŘÍŠKOVÁ 100G" ^ "BON GEISHA 200G"	"MILKA BÍLÁ 100G"	17,65%	100,00%
"MILKA OŘÍŠKOVÁ 100G" ^ "MILKA BÍLÁ 100G"	"BON GEISHA 200G"	17,65%	100,00%

Tato sestava zobrazuje výsledky asociační analýzy - asociační pravidla. Tato pravidla obsahují položky zboží, které se často vyskytují v nákupních košících zákazníků. Asociační pravidlo se skládá z levé a z pravé strany. Na levé straně pravidla jsou položky zboží, které si s určitou pravděpodbností (sloupec podpora) zákazník vždy vloží do košíku. Na pravé straně pravidla jsou uvedeny položky, které si zákazník s určitou pravděpodobností (sloupec spolehlivost) zákazník do košíku ještě přidá.

Obr. 26 Tisková sestava Asociační analýza – výsledkový list

Akční nabídky

Tato tisková sestava zobrazuje jeden z možných případů marketingového využití získaných asociačních pravidel (ukázka viz obr. 27). Položky asociačních pravidel v tomto případě reprezentují akční nabídku, na kterou prodejce poskytuje určitou slevu. Sestava může být také podkladem pro tvorbu dárkových košů, do kterých se umísťují položky, které si zákazníci často žádají společně.

Obr. 27 Tisková sestava Akční nabídky, dárkové koše

8 Závěr

V této práci byly uvedeny některé základní metody dolování znalostí z databází. Byl představen informační systém K2 a některé z jeho klíčových modulů. Na základě uvedených dolovacích úloh a po konzultaci s vedoucím semestrálního projektu bylo zvoleno dolování asociačních pravidel z datového trhu Prodej, který uchovává informace o položkách prodejů, konkrétně analýzu nákupního košíku. Byl navržen modul informačního systému K2 pro získávání znalostí z databází, který byl následně implementován za použití skriptovacího jazyka – K2 skriptu.

Data pro modul dolování znalostí jsou načítána z datového skladu informačního systému K2 a tím tato diplomová práce do jisté míry navazuje na mou bakalářskou prácí (viz [1]).

Nově vytvořený modul získávání znalostí rozšířil stávající možnosti informačního systému K2 a jeho funkčnost byla otestována na testovacím vzorku dat. Demonstrační data obsahovala náhodně vygenerovanou skupinu dokladů prodeje (nákupních košíků). Získané výsledky – asociační pravidla je možné použít pro potřeby marketingu, zejména pro tvorbu nabídek, cílených reklam, dárkových košů a pro optimalizaci rozmístění položek v obchodě. Dolovací proces asociační analýzy obsahuje parametr, kterým lze ovlivnit způsob vyhodnocení frekventovanosti kandidátních množin. Tento parametr povoluje nebo zakazuje testování kandidátní množiny na frekventované podmnožiny. Jeho nastavení je závislé na množství dat zákazníka, ze kterých se získávají asociační pravidla (do jisté míry se jedná o přizpůsobení dolovacího procesu na různé velikosti analyzovaných dat). Zjištění jestli je testovaná podmnožina kandidátní množiny obsažena ve frekventovaných množinách je prováděno pomocí výpočtu hash hodnoty testované podmnožiny a hledání této hodnoty v hash hodnotách frekventovaných množin, což výrazně urychluje celý proces dolování znalostí.

Modul získávání znalostí je v současné době připraven k uvedení na trh a v případě dostatečného zájmu zákazníků bude rozšířen o další funkčnost. Navržená koncepce řešení počítá s tím, že jeden z uživatelů informačního systému bude správcem modulu dolování znalostí, který bude provádět nastavení parametrů a spouštět dolovací proces. Další uživatelé systému pak budou využívat získaná asociační pravidla pro potřeby plánování, řízení a optimalizace. Tento stav je vyhovující v případě nasazení modulu u zákazníků, kteří mají relativně malý tým analytiků. V případě nasazení modulu ve větších společnostech, u kterých je tým analytiků větší, popř. mají několik analytických týmů bude potřeba implementovat rozšíření modulu tak, aby každý uživatel měl k dispozici vlastní sadu parametrů a vlastní množinu výsledků.

Vytvořený modul v současné době obsahuje pouze asociační analýzu nad položkami prodeje. Po předvedení výsledného modulu vedoucímu diplomové práce ze strany zákazníka Mgr. Zdeňku Šarmanovi vznikl návrh na dolování asociačních pravidel z položek výdejů. Získaná asociační pravidla poté budou použitelná pro potřeby optimalizace rozmístění skladu. Pro plné využití potenciálu asociační analýzy bude vhodné implementovat dolování víceúrovňových asociačních

pravidel. Do budoucna se počítá s rozšířením modulu o analýzy dalších typů dolovacích úloh (klasifikace a predikce, shlukování). Tato rozšíření budou implementovat programátoři vývojového oddělení společnosti K2 atmitec s.r.o. z řad studentů Vysoké školy Báňské v Ostravě pod vedením Doc. RNDr. Jany Šarmanové, CSc. jako své diplomové práce v příštím roce. Má diplomová práce se tedy stala prvním počinem v této oblasti a do budoucna tak významně rozšířila možnosti informačního systému K2.

9 Seznam literatury

[1] FIGURA, P. *Export dat z informačního systému K2 společnosti Q.gir s.r.o. pro OLAP*. Brno: VUT, 3. 5. 2005. 56 s. Bakalářská práce.

[2] DUNHAM, M. H. *Data Mining Introductory and Advanced Topics.* 1. vyd. New Jersey: Prentice Hall, 2002. 315 s., ISBN 0-13-088892-3

[3] STRYKA, L. *Modul dolování asociačních pravidel.* Brno: VUT, 29. 5. 2003. *Diplomová práce.*

[4] ZENDULKA, J. *Získávání znalostí z databází, studijní opora.* Brno: VUT, 31. 10. 2006

[5] HAN, J., KAMBER, M. *Data Mining: Concepts and Techniques*. Second Edition. Elsevier Inc., 2006, 770 p., ISBN 1-55860-901-3.

[6] K2 atmitec s.r.o. *Prodej*. Ostrava. Dokument dostupný z WWW: http://www.k2atmitec.cz/produkty/software/moduly.htm?s=prodej [2. 1. 2007]

[7] K2 atmitec s.r.o. *Nákup*. Ostrava. Dokument dostupný z WWW: http://www.k2atmitec.cz/produkty/software/moduly.htm?s=nakup [2. 1. 2007]

[8] K2 atmitec s.r.o. *Marketing*. Ostrava. Dokument dostupný z WWW: http://www.k2atmitec.cz/produkty/software/moduly.htm?s=marketing [2. 1. 2007]

10 Seznam příloh

Příloha 1. Struktury číselníku parametrů

Příloha 2. Stručná uživatelská příručka modulu získávání znalostí z databází

Příloha 3. Podrobný diagram tříd modulu

Příloha 4. CD/DVD, jehož obsahem je:

- nápověda modulu získávání znalostí
- zdrojové kódy modulu a tiskových sestav
- zdrojové texty práce