# KA 3

# Metodické podněty k aplikacím těžení dat a vyhledávání textů

# Obsah



# Seznam obrázků:

Autoři dokumentu: Martin Víta



**ISBN 978-80-88169-15-4**

# Manažerské shrnutí

Data miningové metodiky zastřešují celý proces získávání znalostí z dat a rozčleňují jej do několika po sobě následujících fází. Umožňují „udržet přehled" i v rámci rozsáhlých projektů dobývání znalostí, na němž se podílí větší množství pracovníků (od manažerů přes analytiky až po informatiky) a pracuje se s velkými objemy dat z mnoha heterogenních zdrojů. Implementace této metodiky též eliminuje možnost vzniku určitých typů problémů. V současné době je k dispozici několik metodologií, které jsou neformálně považovány za standardy (podrobněji viz kapitola 2).

Dokument se věnuje možnostem aplikace metodiky CRISP-DM (z angl. Cross Industry Standard Process for Data Mining) na úlohy, řešené v rámci projektu Zefektivnění činnosti TA ČR v oblasti podpory VaVaI a podpory posilování odborných kapacit organizací veřejné správy v oblasti výzkumu, experimentálního vývoje a inovací (dále VaVaI), jmenovitě hledání oponentů, informační podpora oponentů (z hlediska přípravy podkladů – informací o podobných projektech, výsledcích VaVaI aj.), hledání potenciálních duplicit a též analýza struktur v konkrétních oblastech VaVaI.

Metodika CRISP-DM dělí proces do šesti fází, tj.:

1. porozumění problematice (z angl. business understanding),
2. porozumění datům (z angl. data understanding),
3. příprava dat (z angl. data preparation),
4. modelování (z angl. modeling),
5. vyhodnocení výsledků (z angl. evaluation),
6. využití výsledků (z angl. deployment).

V dokumentu je popsán seznam všech konkrétních úkolů/výstupů, které spadají pod jednotlivé fáze, a náhled na obsah jednotlivých fází vzhledem k úlohám.

Uživatelé z řad středního managementu by na základě tohoto dokumentu měli být schopni řídit, resp. organizovat, projekty na implementaci metod zkoumaných v rámci projektu Zefektivnění činnosti TA ČR v oblasti podpory VaVaI a podpory posilování odborných kapacit organizací veřejné správy v oblasti VaVaI (dále Projekt Zefektivnění činností TA ČR) v souladu s metodikou CRISP-DM.

V rámci řešení Projektu byly pořádány prezentace pro zainteresované zaměstnance Kanceláře Technologické agentury České republiky (dále TA ČR), zpětná vazba byla zaznamenávána a formalizována.

# 1. Úvod

Dokument vznikl jako výstup projektu Zefektivnění činností TA ČR, resp. v rámci jeho klíčové aktivity 3 (dále KA 3) - Příprava nových analytických metodik hodnocení VaVaI. Projekt byl financován z Evropského sociálního fondu v rámci Operačního program Lidské zdroje a zaměstnanost.

Hlavním cílem projektu bylo zefektivnění poskytování podpory VaVaI ze strany TA ČR a dalších organizací veřejné správy, posílení odborných kapacit organizací veřejné správy v oblasti VaVaI, posílení chápání významu aplikovaného VaVaI a jeho výsledků pro další rozvoj České republiky a do budoucna i sjednocení způsobů a podmínek poskytování podpory VaVaI.

Cíle projektu Zefektivnění činností TA ČR byly naplňovány sedmi klíčovými aktivitami. Klíčová aktivita KA 3 - Příprava nových analytických metodik hodnocení VaVaI byla zaměřena na podporu posilování odborných kapacit organizací veřejné správy v oblasti VaVaI. Cílem aktivity bylo zefektivnit a zkvalitnit činnosti TA ČR v oblasti analytických služeb jako základní metody zdůvodňování tvorby podkladů pro nastavení a hodnocení podpor. Toho mělo být dosaženo vytvořením nových analytických metodik pro podporu zajišťování hlavních činností a procesů realizovaných TA ČR.

Tento dokument vznikl jako doprovodný materiál ke Studii proveditelnosti pro implementaci modelu k hodnocení nepřímých dopadů výzkumu, experimentálního vývoje a inovací na základě principů modelu StarMetrics™, která je jedním z hlavních výstupů projektu. Byl vypracován na základě analytických zjištění a zkušeností získaných při realizaci KA 3 a to v období od října 2014 do listopadu 2015. Materiál slouží jako jeden z podkladových materiálů pro implementaci kvalitnějších analytických služeb v TA ČR směrem k Evidence Based Policy, jako základní metody zdůvodňování a tvorby podkladů pro nastavení podpor VaVaI.

Hlavním cílem bylo podchytit metodickou aplikaci těžení, konsolidace a vyhledávání dat a textů při realizaci projektu. Metodika by měla poskytnout základní návod pro strukturovaný a systematický přístup k procesu získávání znalostí z dat. Nejedná se pouze o popis samotné metodiky, ale spíše o popis návodu, jak metodicky postupovat při řešení konkrétních úloh. Problematika je tedy názorně předvedena na vybraných úlohách, nicméně ji lze aplikovat na v zásadě na jakýkoliv definovaný problém. Dále jsou v textu používány výrazy data mining a text mining, které autorům připadají v tomto kontextu výstižnější.

# 2. Obecná část

Data miningové metodiky zastřešují celý proces získávání znalostí z dat a rozčleňují jej do několika po sobě následujících fází. Umožňují „udržet přehled“ i v rámci rozsáhlých projektů dobývání znalostí, na němž se podílí větší množství pracovníků (od manažerů přes analytiky až po informatiky) a pracuje se s velkými objemy dat z mnoha heterogenních zdrojů. Implementace této metodiky též eliminuje možnost vzniku některých typů problémů.

V současné době je k dispozici několik metodologií, které jsou neformálně považovány za standardy. Jedná se zejména o následující metodiky:

- CRISP-DM,
- 5A[1],
- SEMMA[2].

Vzhledem k tomu, že dlouhodobě patří mezi nejpoužívanější metody metoda CRISP-DM, byla autory zvolena tato metoda. Lze však předpokládat, že volba jiné z uvedených metodik by vedla k velmi podobným výsledkům.

[1] Metodika se skládá z následujících 5ti kroků: 1. Assess - posouzení potřeb projektu, 2. Access - shromáždění potřebných dat, 3. Analyze - provedení analýz, 4. Act - přeměna znalostí na akční znalosti, 5. Automate - převedení výsledků analýzy do praxe

[2] Metodologie se skládá z následujících 5ti kroků: 1. Sample - vybírání vhodných objektů, 2. Explore - vizuální explorace a redukce dat, 3. Modify -seskupování objektů a hodnot atributů, datové transformace, 4. Model -analýza dat, 5. Assess - porovnání modelů a interpretace

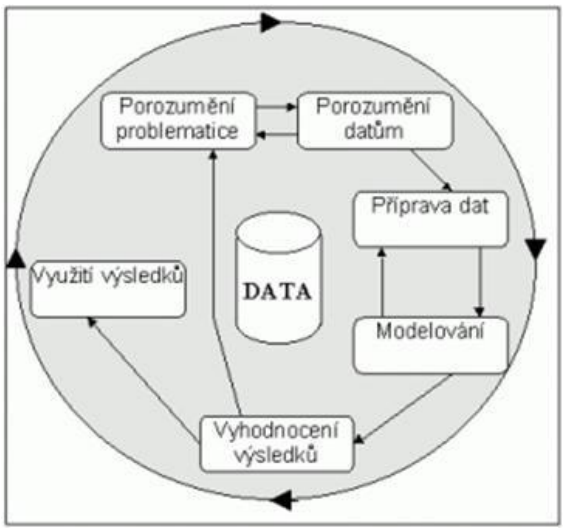
# 3. CRISP-DM

CRISP-DM je metoda, na jejímž vzniku spolupracovaly společnosti NCR Systems Engineering Copenhagen (Spojené státy americké a Dánské království), DaimlerChrysler AG (Spolková republika Německo), SPSS Inc. (Spojené státy americké) and OHRA Verzekeringen en Bank Groep B.V. (Nizozemské království).

Metodika rozděluje životní cyklus data miningového projektu do šesti fází, které na sebe navazují, nicméně nevylučují se (či spíše předpokládají se) „kroky zpět" a iterativní opakování některých sekvencí:

- porozumění problematice,
- porozumění datům,
- příprava dat
- modelování,
- vyhodnocení výsledků,
- využití výsledků.

**Obrázek 1: Fáze CRISP-DM a jejich propojení**




Zdroj: http://www1.osu.cz/studium/dozna/crispdm.htm

# 4. Obecný popis jednotlivých fází metodiky CRISP-DM

## 4.1. Porozumění problematice (z angl. Business Understanding)

Tato fáze obsahuje především formulaci problému manažerským způsobem, popis kontextu a formulaci akceptačních kritérií. Následně jsou problémy převedeny do jasného zadání v termínech data miningu. Součástí této fáze je zpravidla i inventura datových zdrojů.

## 4.2. Porozumění datům (z angl. Data Understanding)

Tato fáze navazuje svým obsahem na inventuru datových zdrojů a je zahájena počátečním sběrem dat. Cílem je získat nadhled nad daty, s nimiž se bude pracovat. Mezi činnosti v této fázi patří zodpovězení otázek týkajících se:

- způsobu získání dat,
- posuzování kvality dat,
- prvního „vhledu“ do dat,
- prvotní selekce relevantních podmnožin záznamů v databázích,
- získání či výpočty různých deskriptivních charakteristik dat (určování četností relevantních atributů, průměrných hodnot, minim, maxim, …).

Vhodnou pomůckou v této fázi může být vizualizace některých dat, např. distribucí hodnot vybraných atributů, které umožňují odhalit některé anomálie a upozornit na případnou nekvalitu vstupních dat.

## 4.3. Příprava Dat (z angl. Data Preparation)

Cílem této fáze je příprava datového souboru či datových souborů, které budou následně zpracovávány analytickými algoritmy (algoritmy dobývání znalostí).

Příprava dat zahrnuje tyto činnosti[3]:

- selekce dat,
- čištění dat,
- transformace dat,
- vytváření dat,
- integrace dat,
- formátování dat.

Je nutné předeslat, že tato fáze bývá často nejpracnější částí řešení celé úlohy.

## 4.4. Modelování (z angl. Modeling)

Tato fáze je vlastní fází dobývání znalostí. Jsou aplikovány algoritmy pro dobývání znalostí na relevantní datasety. Pro řešení dané úlohy typicky existuje více různých metod, které mohou mít odlišné parametry. Tato fáze zahrnuje rovněž i experimentování s použitím uvedených metod/parametrů. Fázi lze opakovat v cyklech, z experimentů může (navíc) vyplynout nutnost

[3] Dobývání znalostí z databází

modifikovat data, což je návrat do předchozí fáze.

Fáze může zahrnovat rovněž testování či ověřování nalezených zákonitostí a vztahů.

### 4.5. Vyhodnocení výsledků (z angl. Evaluation)

V této fázi jsou k dispozici výsledky aplikace algoritmů dobývání znalostí z připravených dat. Výsledky je nutné konfrontovat s manažerskými cíli, které byly formulovány v rámci první fáze: je třeba odpovědět na otázku, zda byly splněny cíle stanovené při zadání.

Na konci této fáze by mělo být přijato rozhodnutí o způsobu využití výsledků (blíže viz Dobývání znalostí z databází).

### 4.6. Využití výsledků (z angl. Deployment)

Životní cyklus data miningového projektu končí nikoliv výběrem a otestováním vhodných modelů (aplikováním data miningových algoritmů na relevantní datasety), nýbrž provedení akcí vedoucích k vlastnímu využívání výsledků: na jedné straně to může být sestavení reportů, závěrečných zpráv atp., na druhé straně může jít o nasazení softwaru, který v sobě implementuje zmiňované modely a je zasazen do procesů dané organizace.

# 5. Poznámky k jednotlivým fázím CRISP-DM

## 5.1. Otázka důležitosti a časové náročnosti jednotlivých fází

Z uvedených popisů jednotlivých fází je zřejmé, že jednotlivé fáze metodiky CRISP-DM mají různou důležitost vzhledem k dosažení cíle a mají též (výrazně) odlišnou časovou náročnost. Jak je uvedeno v materiálu Dobývání znalostí z databází:

- Nejdůležitější je fáze porozumění problému: 75 % významu, 20 % času.
- Časově nejnáročnější je fáze přípravy dat: 75 % času, 20 % významu.
- Vlastní analýzy: 5 % času, 5 %významu.

Uvedené podíly – významnosti/časové náročnosti jsou samozřejmě orientační, nicméně poskytují určité vodítko pro koncipování plánů a předvídání potenciálních problémů.

## 5.2. Přehled příkladů úloh v metodice CRISP-DM v jednotlivých fázích

Pro úplnost je uveden kompletní přehled úloh, které mohou být realizovány v rámci průchodu CRISP-DM metodikou (převzato v původním znění z materiálu Step-by-step data mining guide) a jejich výstupy.

V projektu nebylo předpokládáno, že by v rámci experimentů s implementací této metodiky a obecně data miningových či text miningových metod při řešení reálných úloh TA ČR byly formálně řešeny a zdokumentovány uvedené úlohy. Z důvodu zajištění kompatibility se standardy, které jsou prakticky výlučně v anglickém jazyce, názvy úloh nebyly překládány. Tato část slouží jako pomůcka pro rigorózní zavádění metodiky CRISP-DM.

**Business Understanding**

- **Determine business objectives**

  *Background, Business Objectives, Business Success Criteria*

- **Assess Situation**

  *Inventory of Resources, Requirements, Assumptions and Constraints, Risks and Contingencies, Terminology, Costs and Benefits*

- **Determine Data Mining Goals**

  *Data Mining Goals, Data Mining Success Criteria*

- **Produce Project Plan**

  *Project Plan, Initial Assessment of Tools and Techniques*

**Data Understanding**

- **Collect Initial Data**

  *Initial Data Collection Report*

- **Describe Data**

  Data Description Report

- **Explore Data**

*Data Exploration Report*

- **Verify Data Quality**

  *Data Quality Report*

**Data Preparation**

- **Data Set**

  *Data Set Description*

- **Select Data**

  *Rationale for Inclusion / Exclusion*

- **Clean Data**

  *Data Cleaning Report*

- **Construct Data**

  *Derived Attributes, Generated Records*

- **Integrate Data**

  *Merged Data*

- **Format Data**

  *Reformatted Data*

**Modeling**

- **Select Modeling Technique**

  *Modeling Technique, Modeling Technique Assumptions*

- **Generate Test Design**

  *Test Design*

- **Build Model**

  *Parameter Settings Models, Model Description*

- **Assess Model**

  *Model Assessment, Revised Parameter Settings*

**Evaluation**

- **Evaluate Results**

  *Assessment of Data Mining Results with respect to Business Success Criteria, Approved Models*

- **Review Process**

  *Review of Process*

- **Determine Next Steps**

  *List of Possible Actions Decision*

**Deployment**

- **Plan Deployment**

  *Deployment Plan*

- **Plan Monitoring and Maintenance**

  *Monitoring and Maintenance Plan*

- **Produce Final Report**

  *Final Report, Final Presentation*

- **Review Project**

  *Experience Documentation*

# 6. Praktické aspekty možné implementace CRISP-DM metodiky v prostředí TA ČR

V rámci projektu Zefektivnění činností TA ČR byla prováděna řada experimentů a realizací proof-of-concept některých idejí opírajících se o analýzu dat. Řada z těchto výstupů byla využita pro řešení ad hoc úkolů vycházejících z běžné činnosti TA ČR.

Následující část textu popisuje praktické aspekty možné implementace CRISP-DM metodiky při řešení reálných úloh v prostředí TA ČR. Není tedy přímou aplikací, nýbrž vodítkem, které poskytne přesnou představu, jak při implementaci postupovat. Níže je např. zmíněn požadavek formulace akceptačních kritérií a jsou předloženy jejich ukázky s parametry, avšak není rozhodnuto, jaká kritéria, případně s jakými parametry, budou skutečně použita.

## 6.1. Porozumění problematice

V rámci projektu Zefektivnění činností TA ČR byly vytipovány následující oblasti, ve kterých lze dosáhnout kvalitativního posunu. Jedná se o oblasti:

- Identifikace (potenciálně) duplicitních projektů (dále úloha Duplicity).
- Zkvalitnění doporučování/výběru oponentů k podávaným projektům (dále úloha Oponenti).
- Příprava podkladů pro oponenty (dále úloha Podklady).
- Získání vhledu do struktury entit (projektů, výzkumných pracovníků, institucí, …) v dané oblasti VaVaI (dále úloha Struktury).

### 6.1.1. Úloha Duplicity

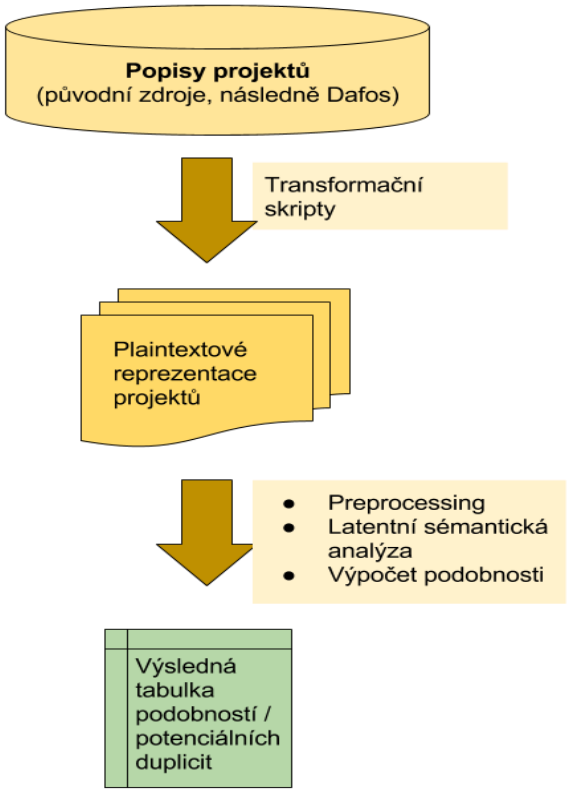
Cílem této úlohy je poskytovat přehledy (dvojic) projektů, které jsou obsahově shodné nebo se v míře překračující jistý práh překrývají s aktuálně či v minulosti řešeným projektem. Na pozadí je snaha vyhnout se duplicitnímu financování stejných či obdobných projektů. Pokud jsou k dispozici informace o míře podobnosti dvojic projektů, lze následně k danému projektu vytvořit seznam ostatních projektů, jejichž podobnost je vyšší než daný práh.

*Současný stav: potenciální duplicity jsou v případě potřeby či podezření vyhledávány zainteresovanými pracovníky (zejména je toto úkolem pro odborné hodnotitele, nyní zajišťované externě). Tito pracovníci však nemají k dispozici vhodný nástroj či možnost získávání reportů obsahujících podněty na prozkoumání podezření na duplicity.*

Formulace v podobě data miningové úlohy: jednotlivé projekty budou reprezentovány textovými soubory (jejich popisy, klíčovými slovy a názvy, případně jinými sadami atributů). Jako potenciální duplicity se označí ty dvojice projektů, u nichž bude podobnost příslušejících textových reprezentací vyšší než předem stanovená hranice.

Kritéria akceptace: expertní posouzení uživateli, zjištění, zda v současné době známé duplicity byly nalezeny, případně zda byla nalezena jejich dostatečná část. Následující schéma znázorňuje základní ideu.

**Obrázek 2: Základní idea duplicit**




### 6.1.2. Úloha Oponenti

Cílem úlohy je vyvinout algoritmus na doporučování oponentů k projektům, který bude výraznějším způsobem akcentovat podobnost obsahu projektu a zaměření potenciálního oponenta.

*Současný stav: aktuálně využívaný algoritmus je založen na určování shody mezi oborovým zatříděním předkládaného projektu a obory potenciálních oponentů (které oponenti zadávají při registraci). Pro tento účel je využívána klasifikace CEP[4]. Následně jsou vyloučeny takové dvojice projekt-oponent, u kterých by docházelo k podjatosti (potenciální oponent je zaměstnancem instituce, která participuje na předkládaném projektu). Dochází však v nemalém procentu případů k tomu, že zaměření oponentů je vzdáleno reálnému obsahu projektu.*

Formulace v podobě data miningové úlohy: oponenti a projekty budou reprezentovány jako textové dokumenty, nový algoritmus bude doporučovat postupně ty oponenty, u nichž je podobnost textové reprezentace s daným projektem nejvyšší s vyloučením případné podjatosti.

Kritéria akceptace: porovnání úspěšnosti přiřazování stávajícím algoritmem a novým. V případě, že procento velmi vhodných či spíše vhodných přiřazení vzroste nad danou mez, bude nový algoritmus dále implementován.

### 6.1.3. Úloha Podklady

Cílem úlohy je poskytovat znalostní podporu oponentům projektů ve smyslu systematického

[4] CEP = Centrální evidence projektů

doporučování relevantních dokumentů (např. popisů projektů, výsledků – publikací, patentů, atd.) majících vztah k oponovanému projektu.

*Současný stav: vyhledávání relevantních dokumentů je věcí iniciativy a schopností jednotlivých oponentů.*

Formulace v podobě data miningové úlohy: analogická jako v případě úlohy Duplicity. Rozdíl je v rozsahu zpracovávaných materiálů, tj. nejen textové reprezentace projektů, ale též výsledků aj.

Kritéria akceptace: expertní posouzení uživateli

#### 6.1.4. Úloha Struktury

Cílem úlohy je systematicky poskytovat informace o strukturách entit v dané oblasti (sítí), zejména z hlediska obsahové podobnosti a vztahů typu spolupráce na stejném projektu, spoluautorství aj. V úloze se jedná o vytipování významných entit v dané síti a určování komunit/clusterů. Součástí této úlohy je rovněž otázka vhodné vizualizace.

*Současný stav: analogické úlohy se řeší, nicméně neopírají se o standardizovaný postup.*

Formulace v podobě data miningové úlohy: struktury budou reprezentovány jako grafy, přičemž entity budou reprezentovány jako vrcholy grafu, relace mezi entitami budou vyjadřovány jako hrany. Význam entit bude modelován pomocí různých druhů centralit, dále budou aplikovány metody na hledání komunit.

Kritéria akceptace: expertní posouzení uživateli

**Inventura dat**

Jako klíčové zdroje dat byly vytipovány:

- Informační systém výzkumu, experimentálního vývoje a inovací, jmenovitě databáze CEP, CEZ a RIV[5].
- Interní informační systémy TA ČR:
  - Informační systém Patriot.
  - Informační systém Beta.

Další zdroje (databáze patentů PATSTAT aj.) mohou být zahrnovány dle potřeby.

**Inventura lidských zdrojů**

Pro realizaci jednotlivých úkolů se předpokládá součinnost programátora/programátorů a analytika/analytiků. Používané technologie a nástroje: SQL, Python, R, Microsoft Excel, Unix shell a jednoúčelové nástroje.

### 6.2. Porozumění datům

Vzhledem k charakteru úloh je zřejmé, že pozornost se soustředí zejména na textová data související s entitami, na které jsou dané úlohy zaměřeny, tj. projekty, výzkumní pracovníci a výsledky VaVaI.

Datové zdroje se rozdělí na základní a doplňkové, přičemž pro potřeby pilotního ověřování konceptů byla využívána především data základní.

[5] CEZ = Centrální evidence výzkumných záměrů, RIV = Rejstřík informací o výsledcích

**Základní data**

- Informační systém Patriot (obsahující data o projektech z veřejných soutěží programů ALFA[6], Omega[7], Centra kompetence[8]) – interní informační systém TA ČR.
- Informační systém Beta – interní informační systém TA ČR, obsahující mj. texty výzkumných potřeb, které byly zadávány jednotlivými resorty.
- Informační systém výzkumu, experimentálního vývoje a inovací, veřejná část –databáze CEP, databáze CEZ a databáze RIV.

**Doplňková data**

- CORDIS[9] – veřejná databáze s webovým rozhraním, relevantní data o projektech s českou účastí.
- Data Národní technické knihovny.
- Data z PATSTAT[10] – vybraná opět pouze část relevantní pro Českou republiku.
- Data z webů kateder vysokých škol České republiky.

**Některé aspekty související se zpracovávanými daty**

Při prvotním seznámení s daty je patrné, že s výjimkou dat z Informačního systému Beta, jsou všechna textová data formálně k dispozici v aglickém jazyce. Při další analýze bylo zjištěno, že některé položky v Informačního systému výzkumu, experimentálního vývoje a inovací (RIV) neodpovídají deklarovanému jazyku, proto je nutné u jednotlivých textových polí, které budou použity v dalším zpracování, provést:

1. detekci jazyka,
2. v případě, že není v anglickém jazyce, tak provést jeho překlad.

Pro účely překladu byl na stroji Cluster zprovozněn systém Moses s jazykovými modely pro překlad z českého do anglického jazyka.

Některá textová data (např. abstrakty článků) v některých případech chybí, nahrazení jinými hodnotami nebo získávání z jiných zdrojů se ukazuje jako problematické, tudíž v praxi bude v malém množství případů pracováno s neúplnými daty.

Rovněž je zřejmé, že budou nastávat problémy s kódováním a používáním bílých znaků (např. v identifikátorech). V rámci Informačního systému výzkumu, experimentálního vývoje a inovací nejsou ve všech případech k dispozici jednoznačné identifikátory osob.

---

[6] Program na podporu aplikovaného výzkumu a experimentálního vývoje ALFA schválený usnesením vlády č. 121 ze dne 8. února 2011 ve znění usnesení vlády č. 669 ze dne 28. srpna 2013

[7] Program na podporu aplikovaného společenskovědního výzkumu a experimentálního vývoje OMEGA schválený usnesením vlády č. 56 ze dne 19. ledna 2011

[8] Program Technologické agentury České republiky na podporu rozvoje dlouhodobé spolupráce ve výzkumu, vývoji a inovacích mezi veřejným a soukromým sektorem Centra kompetence schválený dne 19. ledna 2011 usnesením vlády č. 55

[9] http://cordis.europa.eu/

[10] http://www.epo.org/searching-for-patents/business/patstat.html#tab1

**Vytipování relevantních datasetů**

Pro úlohy Duplicity, Oponenti a Podklady jsou klíčové především položky:

- název entity (projektu, výsledku),
- klíčová slova entity (projektu, výsledku),
- abstrakt entity (projektu, výsledku), či cíl řešení – v rozsahu typicky několika vět/odstavec.

Pro úlohu Struktury jsou klíčové především položky:

- obor (v souladu s klasifikací využívanou v rámci CEP), tj. číselníkové položky,
- identifikátory institucí,
- identifikátory výzkumných pracovníků.

## 6.3. Příprava Dat

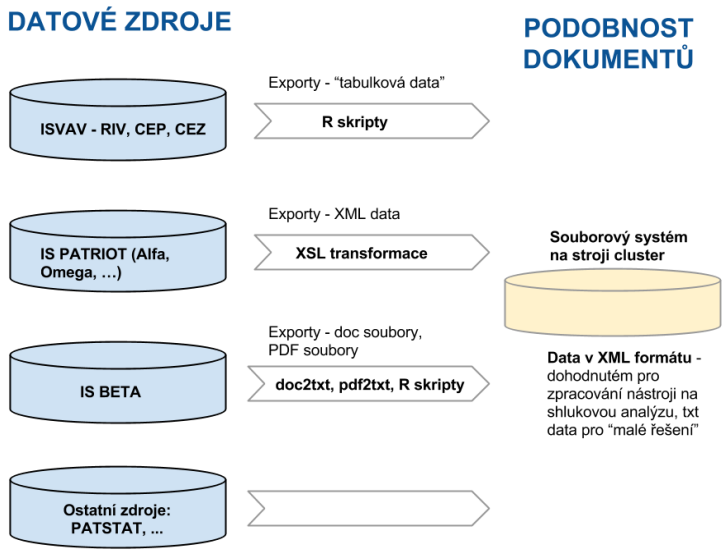
Pro pilotní ověřování konceptu byla z jednotlivých zdrojů získána data z exportů či přímým stahováním jednorázově. V souvislosti s budováním datové základny analytických služeb - DAFOS se předpokládá, že data budou shromažďována v tomto datovém úložišti a podle předem daného rozvrhu budou aktualizována spolu s kontrolami integrity.

**Současný stav / proces transformace dat**

Tento stav je naznačen na schématu níže.

Z výsledných XML jsou v případě potřeby (zejména „malé zpracování“ – viz následující fáze) připravovány plaintextové soubory.

**Obrázek 3: Současný stav / proces transformace dat**




Poznámka: ISVAV = Informační systém výzkumu, experimentálního vývoje a inovací

## 6.4. Modelování

Úlohy Duplicity, Oponenti a Podklady jsou založeny na určování podobnosti textů, přesněji textových

reprezentacích relevantních entit (oponentů, projektů). Jádrem řešení je volba dalšího zpracování textových souborů a výpočet podobnosti.

Z hlediska IT vybavení budou rozlišeny dvě situace:

- kolekce dokumentů čítá méně než tisíce dokumentů: v takovém případě bude řešena úloha na běžném PC se systémem R,
- kolekce dokumentů je na úrovni vyšších tisíců (v případě zpracování všech dokumentů v RIV a vybraných dokumentů databáze PATSTAT se jedná o řádově 1 mil. dokumentů). Tato situace bude řešena s využitím vybavení, popsaném v rámci materiálu Metodika postupu shlukové analýzy.

**Základní princip výpočtu podobnosti pro úlohy Duplicity, Oponenti a Podklady**

Při preprocessingu textů z kolekce se provádí běžná sekvence procedur (případně zvolená podmnožina):

- převod na malá písmena,
- odstranění diakritiky,
- odstranění číslic,
- odstranění stop slov, případně slov kratších než tři znaky.

Cílem preprocessingu je odstranit nerelevantní řetězce/slova a zabránit rozlišování řetězců, které se liší pouze velikostí písmen. Pro textové dokumenty je zvolena vektorová reprezentace (pro bag-of-words), v rámci této kolekce je použito *tf-idf* váhování.

Jako míra podobnosti dokumentů je zvolena cosinová podobnost. Při dalších experimentech lze volit i jiné podobnostní metriky.

Poznámky k jazykovým verzím dokumentů: pro zjednodušení zpracování byl zvolen anglický jazyk, dokumenty v jiném jazyce (např. český jazyk) jsou do anglického jazyka automaticky překládány pomocí implementace systému MOSES na stroji Cluster. V případě výpočtu podobnosti nad dokumenty výzkumných potřeb z Programu veřejných zakázek ve výzkumu, experimentálním vývoji a inovacích pro potřeby státní správy Beta (usnesení vlády č. 54 ze dne 19. ledna 2011 změněný usnesením vlády č. 75 ze dne 31. ledna 2013) jsou však používány dokumenty v českém jazyce.

Pro úlohy Duplicity a Podklady byly do výstupu vybrány takové dvojice projektů, jejichž podobnost je vyšší než daný threshold. V případě úlohy Oponenti bylo voleno 6 (či jiný požadovaný počet) nejpodobnějších oponentů k danému projektu.

**Úloha Struktury**

Entity (např. výzkumní pracovníci) jsou reprezentovány jako vrcholy grafu, hrany reprezentují vztahy mezi entitami (např. spolupráce na projektu).

Pro detekci komunit byl vybrán algoritmus WalkTrap, pro modelování významnosti byla vybrána eigenvector centrality. V rámci dalších experimentů byly využívány jiné typy centrality (closeness, betweenness aj.).

Grafová data, popisující strukturu, byla vizualizována s využitím standardních prostředků systému R. Tloušťka hran odpovídá síle vazby, velikost vrcholu významnosti, komunity jsou podbarvovány.

**Evaluace**

Úlohy Duplicity a Podklady spadají do oblasti klasického information retrieval. Lze tedy použít klasických metrik, které se v této oblasti používají. Jedná se zejména o analýzu matice záměn (z angl. confusion matrix) čili zjišťování těchto čtyř parametrů (pro ilustraci je provedena na úloze Duplicity):

- true positives – počet skutečných duplicit, které byly uvedeným postupem označeny jako duplicity,
- true negatives – počet „neduplicit“, které byly uvedeným postupem označeny jako neduplicitní,
- false positives – počet neduplicit, které jsou (nesprávně) označeny jako duplicity,
- false negatives – počet duplicit, které jsou (nesprávně) označeny jako neduplicitní.

Je zřejmé, že náročnost zjišťování jednotlivých parametrů je různá. Počet false negatives je v kontextu uvedených úloh obtížné zjistit, neboť by to reálně znamenalo mít přehled o všech duplicitách, které jsou v daném posuzovaném souboru.

V případě zbylých dvou úloh byla kritéria volena dle charakteru úlohy.

**Úloha Duplicity**

Výstupy byly evaluovány uživateli, především z Oddělení podpory hodnocení projektů TA ČR. Klíčovým parametrem je počet neduplicit, které jsou nesprávně označeny jako duplicity – velký podíl nerelevantních dokumentů snižuje efektivitu následného zpracování dokumentů (nutnost další „lidské“ práce s výsledky).

Na omezených souborech (pro účely testování) byl zkoumán i podíl duplicit, které byly nesprávně označeny jako neduplicitní. Zvažovalo se i o vytvoření kolekce „fiktivních“ duplicitních projektů, které by pomohly lépe charakterizovat problematické chování zvoleného algoritmu.

**Úloha Oponenti**

Výsledky přiřazování byly evaluovány oproti výsledkům stávajícího systému. Na vybrané reálné sadě projektů se hodnotili oponenti vybraní jednotlivými systémy. Hodnocení mělo charakter Likertových škál (rozhodně nevhodný, spíše nevhodný, neutrální, spíše vhodný, velmi vhodný) pro každé přiřazení oponenta projektu. Na závěr se agregují data pro jednotlivé algoritmy přes všechny posuzované projekty a oponenty.

Klíčovým rozhodovacím parametrem je podíl přiřazení s kladnou polaritou (spíše/velmi vhodný). V případě, že tento parametr přesáhne danou mez, resp. zlepšení oproti stávajícímu algoritmu o více než dané procento, je přistoupeno k implementaci nového algoritmu.

**Úloha Podklady**

Výstupy jsou evaluovány uživateli, kteří připravují materiály pro oponenty. Sledovaná kritéria jsou totožná s úlohou Duplicity.

**Úloha Struktury**

Výstupy byly evaluovány uživateli s příslušným odborným zázemím. Posuzoval se podíl výstupů, které po expertním zhodnocení nedávají relevantní výsledky. V případě situace, kdy tento podíl přesáhne předem danou mezní hodnotu, přikročí se k využití jiných metod (např. jiných druhů centralit aj.).

## 6.5. Využití výsledků

Vzhledem k tomu, že řešení jednotlivých úloh (Duplicity, Oponenti, …) mají formálně podobu tabulek, případně tabulek doplněných o další data (např. seznam dvojic potenciálně duplicitních projektů, který je doplněn o jejich krátké popisy, identifikační údaje či metadata), spočívá fáze Deployment především v přípravě (formátování, konvertování) těchto výstupů a jejich distribuci relevantním zaměstnancům (např. Oddělení podpory hodnocení projektů TA ČR, analytickým pracovníkům aj.).

V rámci zmíněné fáze je/bude nutné navrhnout proces zpětné vazby od uživatelů z hlediska životního cyklu těchto data miningových úloh – zejména pro účely zlepšování kvality a uživatelského komfortu.

V budoucnu lze automatizovat tvorbu některých reportů v návaznosti na požadavky uživatelů a integraci do (uživatelského rozhraní) připravovaného informačního systému.

**Úloha Duplicity**

Výstupy měly podobu tabulek a popisů projektů distribuovaných k zaměstnancům emailově či sdílením na GoogleDisku.

**Úloha Oponenti**

Výstupy měly v pilotní verzi podobu tabulek, která byla předávána ke zhodnocení vybraným členům Oddělení podpory hodnocení projektů. V budoucnu, po splnění akceptačních podmínek, se předpokládá k přistoupení k integraci informačního systému.

**Úloha Podklady**

Seznamy relevantních dokumentů, které budou případně doplněny plnými texty, budou předávány těm, kteří mají ve své agendě přípravu materiálů pro oponenty a komunikaci s nimi. Lze předpokládat, že při přípravě materiálů bude docházet k manuálním korekcím (vyřazování nerelevantních dokumentů).

**Úloha Struktury**

Výstup má podobu kolekce dokumentů obsahující typicky seznamy entit v dané oblasti, kvantitativní data (např. hodnoty centralit v rámci grafové struktury) a vizualizace. Tyto informace budou distribuovány opět emailem případně sdílením na GoogleDisku.

# 7. Seznam použitých zkratek

| | |
|---|---|
| **Metodika CRISP-DM** | z angl. Cross Industry Standard Process for Data Mining |
| **Projekt Zefektivnění činností TA ČR** | Zefektivnění činnosti TA ČR v oblasti podpory VaVaI a podpory posilování odborných kapacit organizací veřejné správy v oblasti VaVaI |
| **TA ČR** | Technologická agentura České republiky |
| **VaVaI** | Výzkum, experimentální vývoj a inovace |

# 8. Seznam použitých zdrojů

**Internetové zdroje**

- https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
- http://www1.osu.cz/studium/dozna/crispdm.htm
- http://cordis.europa.eu/
- http://www.epo.org/searching-for-patents/business/patstat.html#tab1

**Ostatní**

- Berka, P. Dobývání znalostí z databází. Academia [online]. In: Portál Databáze knih. 2003. ISBN:80-200-1062-9 [http://www.databazeknih.cz/knihy/dobyvani-znalosti-z-databazi-78521]
- Chapman, P. et all. CRISP-DM 1.0, Step-by-step data mining guide. In: Portál The modeling agency. SPSS, Inc. 2000 [https://www.the-modeling-agency.com/crisp-dm.pdf]
- Program na podporu aplikovaného výzkumu a experimentálního vývoje ALFA schválen usnesením vlády č. 121 ze dne 8. února 2010 [online]. In: Portál Technologické agentury ČR [https://www.tacr.cz/index.php/cz/programy/program-alfa.html]
- Změna programu na podporu aplikovaného výzkumu a experimentálního vývoje ALFA schválena usnesením vlády č. 669 ze dne 28. srpna 2013 [online]. In: Portál Technologické agentury ČR [https://www.tacr.cz/index.php/cz/programy/program-alfa.html]
- Program Technologické agentury České republiky na podporu rozvoje dlouhodobé spolupráce ve výzkumu, vývoji a inovacích mezi veřejným a soukromým sektorem Centra kompetence schválený usnesením vlády č. 55 ze dne 19. ledna 2011 [online]. In: Portál Technologické agentury ČR [https://www.tacr.cz/index.php/cz/programy/centra-kompetence.html]
- Program veřejných zakázek ve výzkumu, experimentálním vývoji a inovacích pro potřeby státní správy BETA schválený usnesením vlády č. 54 ze dne 19. ledna 2011 [online]. In: Portál TA ČR [https://www.tacr.cz/index.php/cz/programy/program-beta.html]

Martin Víta

KA 3

Metodické podněty k aplikacím těžení dat a vyhledávání textů

Vydala: Technologická agentura ČR, Evropská 1692/37, 160 00 Praha 6

http://www.tacr.cz

Praha 2016

1. vydání



**ISBN 978-80-88169-15-4**