FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

ZADÁNÍ DIPLOMOVÉ PRÁCE

Název:	Měření kvality kvantitativních zdrojů pro automatizovaný forecasting nových technologií
Student:	Bc. Michal Režnický
Vedoucí:	doc. RNDr. Ing. Marcel Jiřina, Ph.D.
Studijní program:	Informatika
Studijní obor:	Webové a softwarové inženýrství
Katedra:	Katedra softwarového inženýrství
Platnost zadání:	Do konce letního semestru 2018/19

Pokyny pro vypracování

Forecasting se zabývá předpovědí budoucího vývoje. S přibývajícími daty lze forecasting automatizovat. Cílem práce je prozkoumat oblast vhodných kvantitativních zdrojů pro automatizovaný forecasting pro predikci vývoje nových technologií.
1) Formulujte pojem forecastingu. Analyzujte možnosti automatizovaného forecastingu nových technologií.
2) Analyzujte ekonomické přínosy/dopady automatického forecastingu při strategickém plánování budoucího směřování podniku a zavádění nových technologií.
3) Identifikujte a analyzujte kvantitativní zdroje potenciálně vhodné pro automatizovaný forecasting se zaměřením na predikci nových technologií.
4) Navrhněte metriky pro měření kvality kvantitativních zdrojů pro automatizovaný forecasting pro predikci nových technologií a aplikujte je na identifikované zdroje.
5) Zvolte výpočetní model pro ohodnocení navržených metrik váhami, implementujte jej a stanovte příslušné váhy.
6) Diskutujte dosažené výsledky, zejména možné dopady na ekonomický přínos.

Seznam odborné literatury

Dodá vedoucí práce.

Ing. Michal Valenta, Ph.D.
vedoucí katedry

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
děkan

V Praze dne 2. února 2018

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra Softwarového inženýrství

Diplomová práce

Měření kvality kvantitativních zdrojů pro automatizovaný forecasting nových technologií

Bc. Michal Režnický

Vedoucí práce: doc. RNDr. Ing. Marcel Jiřina, Ph.D.

27. dubna 2018

Poděkování

Děkuji svému vedoucímu doc. Jiřinovi za inspirativní a motivující konzultace a velký zájem o mou práci. Děkuji své rodině a blízkým, kteří ve mě věřili a podporovali mě.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou, a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla, a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu), licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 27. dubna 2018 .....

Odkaz na tuto práci

Režnický, Michal. *Měření kvality kvantitativních zdrojů pro automatizovaný forecasting nových technologií.* Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2018.

Abstrakt

Tato diplomová práce se zabývá měřením kvality kvantitativních zdrojů pro automatizovaný forecasting nových technologií. Analytická část práce pokrývá různé typy forecastingu jak po technické, tak po ekonomické stránce. Stěžejní částí práce je návrh vlastních metrik pro měření kvality kvantitativních zdrojů pro automatizovaný forecasting nových technologií. Návrh je proveden na základě analýzy měření kvality zdrojů v jiných oborech. Dle navržených metrik jsou zdroje ohodnoceny. Nakonec jsou metrikám přiděleny váhy důležitosti a vypočteny kvality konkrétních zdrojů.

Klíčová slova Forecasting nových technologií, forecasting, automatizovaný forecasting, kvalita dat

Abstract

This master thesis is dealing with measurement of quality of quantitative sources for automated forecasting of emerging technologies. Analytical part of this thesis covers different forecasting fields from the technical point of view as well as economical. Fundamental part of this thesis is design of original metrics for measuring the quality of quantitative sources for automated forecasting of emerging technologies. The basis for designed metrics lays in the analysis of methods of measuring sources quality in other fields. The sources are rated by the rules of designed metrics. The metrics are given weights by importance and several sources quality are calculated.

Keywords Forecasting emerging technologies, forecasting, automated forecasting, data quality

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Předvídání budoucnosti je důležitým aspektem v každodenním rozhodování všech lidí. Velkou roli hraje nejen pro lidi, ale i pro podniky, jež na základě konkrétních očekávání budoucího vývoje řídí své aktivity. Forecasting je obor, který se snaží různými metodami provádět předvídání budoucnosti. Jeho cílem je zpřesňovat předpovědi, a pomáhat tak lidem i firmám.

Některé typy forecastingu lze automatizovat, aby jej nemuseli provádět pouze lidé. Tím se ušetří mnoho nákladů a zefektivní celý proces.

Forecasting nových technologií předvídá růst zájmu o vznikající technologie, které byly doposud známé pouze úzkému okruhu lidí, většinou těch, kteří se podíleli na jejím vzniku. Na základě znalosti vznikajících technologií mohou firmy podnikat strategické investice a rozhodnutí, které je mohou výrazně posílit.

Celý proces lze automatizovat, a to se nazývá automatizovaný forecasting nových technologií. Místo, odkud bere data pro své výpočty, se nazývá zdroj dat. A právě výběru zdroje dat pro automatizovaný forecasting nových technologií se bude věnovat tato práce.

Cíl práce

Práce má v první části zpracovat analýzu oboru forecastingu a jeho variant – automatizovaného forecastingu, forecastingu nových technologií a automatizovaného forecastingu nových technologií, a to z pohledu technického i ekonomického. Cílem je zejména zaměřit se na automatizovaný forecasting spojený s forecastingem nových technologií.

Následuje druhá část, jejíž úkolem je vytvořit rámec pro hodnocení kvality zdrojů pro použití v automatizovaném forecastingu nových technologií. Jako první je nutné identifikovat potenciální kvantitativní zdroje vhodné pro použití v automatizovaném forecastingu nových technologií. Na základě analýzy kvality zdrojů budou navrženy metriky pro měření kvality zdrojů, které budou

na identifikované zdroje aplikovány. Toto je stěžejní část práce, bude se jednat o původní dílo autora.

Po určení vah metrik dle zvoleného modelu budou na základě ohodnocení vhodných zdrojů vypočteny kvality těchto zdrojů. Na závěr budou diskutovány výstupy práce s důrazem na ekonomický přínos práce.

Kapitola 1

Vymezení pojmu forecasting

1.1 Úvod do forecastingu

Forecasting, česky prognózování, prognostika či predikování, je činnost, která se zabývá odhadem budoucnosti. Jedná se o odhad v nejširším slova smyslu, prognózovat se dají všechny myslitelné jevy, ať už jde o vývoj počasí, směřování nějaké firmy, ukazatele výkonnosti, vítěze budoucích voleb nebo hodnotu akcií. Prognózovat se dá ale také vývoj počítačů – jaké budou v budoucnu velikosti pamětí, rychlosti procesorů. Jako zdroj dat se využívají například preference voličů, směry větru a aktuální parametry počítačových komponent.

Předmětem forecastingu ale také mohou být nečíselné veličiny. Jaké by to bylo odhadnout, co bude aktuální za deset let v oblasti technologií? Jestli se již nyní rodí 4D tisk nebo něco více než samořiditelná auta, jaké jsou trendy v oblasti obnovitelných energií? Jsou to novinky, kterým se nyní věnují pouze výzkumníci a do produkce či k běžným uživatelům se dostanou třeba až za dekádu.

To je velmi zajímavé téma a většina aktuálních konvenčních technik forecastingu na ně nedosáhne. V dnešní době se forecasting pro podniky točí především kolem předvídání čísel, číselných řad. Ty jsou využitelné například na burzách všeho druhu, predikují vývoj cenných papírů, kurzu měn nebo HDP zemí. Dále v predikci finančních ukazatelů – jak si firma povede v následujícím roce, kolik bude mít zakázek, kolik vydělá, jak má investovat svůj rozpočet co nejefektivnějším způsobem. Z číselných řad se dají trendy extrahovat poměrně dobře a toto území je zkoumáno již několik desetiletí. V dnešní době je prognóza číselných řad tak běžná, že ji zvládá i jeden z nejrozšířenějších kancelářských softwarů, Microsoft Excel, již více než 10 let.

V následujících kapitolách bude představeno dělení a struktura současných technik forecastingu. Také bude definováno, jak tato práce chápe technologický forecasting, což bude důležité pro porozumění směru, kterým se práce vydá.

1.2 Definice forecastingu

Existuje mnoho definic forecastingu. Liší se zejména širokými možnostmi aplikace, z nichž některé jsou pro ilustraci naznačeny v předchozí sekci. V následujících odstavcích je popsáno několik obecných definic, které by měly dobře vystihovat podstatu tohoto oboru.

První definice se nachází v jedné z výzkumných prací od MIT, konkrétně v článku „Technological Forecasting - a Review“ [1]. Tvrdí, že forecasting je záměrná a systematická snaha o pochopení a uchycení potenciálního směru, hodnot, charakteristik a efektů zkoumaného jevu, více detailů o tomto výzkumu se nachází v části 1.4.3.2

Nejpřístupnější definice, tedy ta, která je zájemci o zjištění významu slova forecasting nalezena mezi prvními ve vyhledávačích, se nachází na anglické stránce Investopedia. Ta samozřejmě není ověřeným zdrojem informací a nehodí se pro výzkumné účely, protože není recenzovaná odbornou veřejností. Tvrdí, že forecasting je užití historických dat k odvození směru budoucích trendů. Dodává důležitou poznámku, že čím dále se nahlíží do budoucnosti, tím jsou predikce méně přesné [2].

Třetí zde zmíněná definice je oficiální význam slova „forecast“ ze slovníku Mirriam-Webster, jež je přes třicet let součásti prestižní Encyclopædia Britannica, a tudíž se jedná o etablovaný zdroj slovníkových definic. Dle něj provádět forecasting znamená predikovat nějakou událost nebo stav v budoucnosti na základě výsledku studia a analýzy vhodných dat. Vhodná data jsou taková, která se k predikované události vztahují [3].

Není třeba zavádět vlastní definici, jelikož ta poslední, třetí, nejlépe zachycuje esenci forecastingu tak, jak je v této práci chápan. Práce se na ní bude také odkazovat.

1.3 Členění forecastingu

Jak je naznačeno v úvodu, forecasting je velice obecný pojem a lze ho rozčlenit na podskupiny popsané v této sekci. K lepšímu uchopení tohoto pojmu bude nutné seznámit se s jeho způsoby aplikace. Už bylo řečeno, co predikování znamená, ale jak jej použít, jak jej převést do praxe, jak vůbec predikovat? Na to nám odpoví následující přehled metod implementace forecastingu.

Správnou metodu není jednoduché vybrat. Osoba plánující vytvořit prognózu by se měla sama sebe dotázat, k jakému účelu ji využije, jaké jsou proměnné, které se snaží předvídat a také jak zásadní roli hraje minulost neboli historická data v souvislosti s předvídaným vývojem. A zejména jaká data vůbec má k dispozici.

Základním rozdělujícím prvkem je typ vstupních dat. Dle něj jsou metody aplikace forecastingu děleny na kvalitativní a kvantitativní. Tak budou také rozděleny v této práci.

Základním krokem pro pochopení tohoto rozdělení bude vysvětlení významu kvalita dat a kvantita dat. Pro ilustraci bude užit příklad datové entity – uživatelský profil.

Kvalitativní metody si zakládají na množství informací o jedné entitě, tedy do jakých podrobností jsou data rozpracována, hlavní roli hraje kvalita. U datové entity uživatelský profil by to bylo mnoho osobních informací a preferencí, například až 50 parametrů. Díky tomu by šel jedinec, kterému profil patří, velmi dobře identifikovat a zpracovatel dat by získat velmi přesnou představu, o koho jde. Na druhou stranu těchto kvalitně vyplněných uživatelských profilů by bylo dispozici pouze několik desítek, z těchto dat nelze příliš dobře odvodit průměrné hodnoty parametrů pro všechny jedince. Pro kvalitativní forecasting je důležité mít přesná data, tato jsou sbírána často od expertů na danou doménu. Tím pádem bývají kvalitativní data často subjektivně zabarvená.

Oproti tomu kvantitativní využívá množství informací je lhostejné, pokud jsou některá data mírně zkreslená, ta se v tak velkém objemu ztratí. U datové entity uživatelský profil by mohlo být k dispozici méně parametrů, například 10, ale samotných uživatelských profilů by byly tisíce. Samotní jedinci vlastnící profily by nebyli lehce identifikovatelní, ale bylo by snadné odvodit průměrné hodnoty parametrů celé skupiny jedinců. Kvalitativní metody lze dobře automatizovat, často se jedná o statistické výpočty. Vstupem jsou často exaktní data, která lze dobře počítačově reprezentovat a provádět s nimi operace.

Krátce shrnuto, kvalitativní metody se opírají o detailní znalost domény a na základě toho provádí předpovědi, kdežto kvantitativní staví na množství dat, které bylo nashromážděné. Kvantitativní data mohou být dokonce anonymizována tak, že ten, kdo provádí prognózu, nemusí vědět, čeho se týkají a přesto z nich dokáže vypočítat uspokojivou předpověď.

Následující dělení forecastingu dle způsobu aplikace je odvozeno z několika publikací věnujících se forecastingu. Mnoho zdrojů vidí rozdělení různě, například není jasně daná hranice mezi statistickými metodami a časovými řadami. Oboje spolu souvisí, zároveň v nalezené literatuře nepanuje všeobecný konsenzus, jaké metody kam spadají. Následující rozdělení je popsáno na základě průzkumu knihy „Statistical Methods for Forecasting“ [4] a přednášky „Time Series Forecasting Methods“ [5].

1.3.1 Kvantitativní metody

Kvantitativní forecasting používá historická data, pro firmy to mohou být například uplynulé objemy prodejů, produkce nebo finanční reporty.

Jednoduchým příkladem je firma vyrábějící jízdní kola. Stačí jim shromáždit počty prodejů jednotlivých modelů kol za poslední měsíce a příslušné zisky. Nahlédnutím na tato data dokáží promítnout očekávané prodeje a zisky. Stejně jako ve zmíněném případě je vhodné metodu užít právě tehdy, když očekáváme, že se trend bude držet nebo opakovat i v budoucnosti.

Jde v podstatě o vytvoření funkce, kde jsou budoucí hodnoty závislé na historických hodnotách.

Pro výpočet kvantitativních metod je často užita statistika. Většina zmíněných metod používá k výpočtu postupy statistiky. Vstupní data bývají především v číselné formě.

1.3.1.1 Časové řady

Predikce budoucích stavů vycházejí z číselných dat dostupných v minulosti. Jsou velice vhodné pro krátkodobé až střednědobé plánovaní. Mají široké využití v oborech jako bankovnictví (vývoj kurzu, HDP, cen) a podnikání (predikce nákupů, prodejů, počtu uživatelů).

Naive approach

Jednou z nejjednodušších metod je naivní metoda (Naive approach). Ta je založena na dostatečně dlouhé číselné řadě, z níž vybere poslední číslo. Tato metoda je většinou efektivní u predikcí finančních sérií, kde jsou jednotlivé výkyvy velice těžko odhadnutelné.

$$forecast = y_t \tag{1.1}$$

Proměnná y obsahuje data z minulosti, mohou to být například hodnoty ukazatele v jednotlivých měsících, a t bude v tom případě počet měsíců, tedy čas. Samotné y_t je hodnota dat z minulosti v měsíci t.

Tento způsob má mnoho úprav, tzv. Seasonal Naive approach by do proměnné y přiřazoval vždy konkrétní měsíc, třeba leden. V některých byznys případech jsou totiž stavy ukazatelů korelované s konkrétním měsícem (například nižší návštěva restauračních zařízení v lednu a únoru).

Average approach

Další intuitivní metodou je metoda průměru (Average approach). Zde se jednoduše vezme průměr minulých hodnot a prohlásí se za predikci.

$$forecast = \frac{1}{t}\sum_{i=1}^{t} y_i \tag{1.2}$$

Samotná proměnná t znamená celkový počet měsíců, přičemž y_t je i nadále hodnota dat z minulosti v měsíci t.

Seasonal Average approach znovu do proměnné y přiřadí konkrétní měsíc. Tady je vidět, že přidání sezóny do predikce je celkem populární a opravdu dává smysl, jak je popsáno u Seasonal Naive approach v předchozí části.

Moving average (MA)

Model klouzavých průměrů je, jak název napovídá, mírně upravená metoda průměru. Hlavní idea této metody je, že se průměry pohybují s časem, tedy že některé starší již nebudou brány v potaz. Na číselné řadě $y_1, ..., y_n, ..., y_t$ je definováno symetrické pohyblivé okno o velikosti $2m + 1$, to se s časem t pohybuje. Výpočet se provádí dle následujícího vzorce:

$$forecast = \sum_{i=-m}^{m} \frac{y_{n+i}}{(2m+1)} \tag{1.3}$$

Mírnou úpravou je přidání vah, se kterými lze časové body, například měsíce, ohodnotit. Strategie ohodnocení závistí na tom, kdo provádí výpočet.

Na vážený klouzavý průměr navazuje **Exponential smoothing**, česky exponenciální vyhlazování, které určuje časovým bodům předem definované váhy. A to tak, že čím je časový bod starší, tím bude mít ve výsledku nižší váhu. Existuje několik variant exponenciálního vyhlazování dle strategie přiřazení vah.

Dalšími pokročilými modely pracujícími s průměrem jsou Autoregressive Moving Average model (ARMA) a Autoregressive Integrated Moving Average model (ARIMA). Oba dva mají i své adaptace s přidáním sezón – Seasonal Autoregressive Moving Average model (SARMA) a Seasonal Autoregressive Integrated Moving Average model (SARIMA) [4].

Drift method

Metoda driftů identifikuje „drift“, tedy změnu mezi historickými daty, a reflektuje jí ve své předpovědi. Ekvivalentem je nakreslení rovné čáry od začátku dat až po konec a její protažení do budoucna, jak je vidět z upraveného vzorce na pravé straně.

$$forecast = y_t + h(\frac{\sum_{i=2}^{t} y_i - y_{i-1}}{t-1}) = y_t + h(\frac{y_t - y_1}{t-1}) \tag{1.4}$$

Kde h je velikost posunu do budoucna, tedy o kolik časových úseků dopředu chceme předpověď vypočítat. Forecast je tedy pro čas $t + h$.

Další metody časových řad

Modelů pro časové řady je opravdu mnoho, pro úplnost je ještě zmíněna **Extrapolation**, česky extrapolace. Jedná se o metodu, kdy je odhad budoucí hodnoty proveden na základě jejího vztahu s jinou hodnotou.

Pro získání představy o predikcích hodnot z časových řad je dosavadní výčet postačující.

1.3.1.2 Bibliometrie

Bibliometrie je obor, který se zabývá kvantitativní analýzou dokumentů s vědeckými poznatky. Konkrétně jde o analýzu patentů, abstraktů, odborných

prací a knih. V dalším kroku je možné vytvářet clustery ze získaných dat a ty procházet a těžit. Jedná se o data mining, konkrétně text mining.

Motivace pro vznik této práce je právě aplikace metod bibliometrie, jak bude vysvětleno v sekci 1.4, kde se nachází i konkrétní popis užití bibliometrie.

1.3.2 Kvalitativní metody

Jak již bylo řečeno u dělení forecastingu, kvalitativní metody jsou více subjektivní než kvantitativní a také spoléhají na detailnější znalost domény. Vstupní data bývají především v textové formě. Data jsou získávána těžením (data mining) nebo vyplňováním dotazníků. Predikce jsou často založené na tzv. expertních odhadech. Expert je někdo, kdo dané doméně rozumí a má dostatek informací k tomu, aby dokázal předvídat budoucí vývoj.

1.3.2.1 Delfská metoda

Původ názvu pochází ze starého Řecka z města Delfy (Delphi), kdy věštkyně, tedy experti, věštily budoucnost za použití omamných výparů.

V dnešní době je delfská metoda prováděna formou dotazování expertů, kteří nezávisle na sobě odpovídají na stejné otázky. Tento postup může být jednokolový nebo vícekolový, záleží na tom, jaké informace se organizátor snaží zjistit. Pokud je vícekolový, tak jsou po každém kole anonymně zveřejněny odpovědi a experti mají možnost upravit svou výpověď. Po několika kolech se na základě nějakého pravidla (počet kol, průnik názorů) organizátor spokojí s výsledkem.

1.3.2.2 Brainstorming

Jedná se o skupinové kreativní techniky a liší se pouze formou zápisu závěrů. Průběh je intuitivní, je svoláno množství expertů a ti diskutují, případně sepisují nápady, jak by mohl vypadat budoucí vývoj. Znovu je potřeba moderátor, který diskuzi bude organizovat.

Jako část Brainstormingu lze uchopit Participativní metodu, která umožňuje sdílet osobní zkušenosti účastníků a tím prohlubovat jejich znalosti.

1.3.2.3 Sestavování scénářů

Skupina expertů se sejde nad vymýšlením alternativních scénářů, ze kterých se snaží vyvodit pravděpodobné závěry. Odhadují, co by se za konkrétních podmínek a nastavení proměnných mohlo v budoucnu stát. Tato metoda funguje dobře při předpovídání faktů, na které jsou statistické metody krátké, a to například příchodu nového konkurenčního výrobku na trh. Experti odhadují, jaký by mohl mít nový výrobek dopad na jejich firmu a prodej jejich výrobků.

1.3.3 Metody umělé inteligence

Na pomezí kvantitativních a kvalitativních přístupů stojí umělá inteligence. Sama o sobě provádí zadané výpočty, spíše je to tedy implementace nějakého složitého modelu, například neuronové sítě. Ty jsou využívané nejčastěji, protože dobře operují s velkým množstvím dat.

Oproti tomu ale stojí data mining, který může pracovat s obojím – jak kvalitou, tak kvantitou, jak je zmíněno u v podsekci Bibliometrie 1.3.1.2. Extrahování číselných řad z různých zdrojů je jedna možnost, druhou ale je extrahování expertních dat, například z knížek a vědeckých studií. Příkladem takového systému je IBM Watson, konkrétně jeho odnož Health [6]. Ten „nastudoval“ nespočet vědeckých článků a reportů o zdraví a během let se etabloval jako opravdu praktický pomocník lékařů, kteří si za ním chodí pro radu.

1.4 Forecasting nových technologií

V této sekci bude vymezen termín forecastingu nových technologií tak, jak bude uvažován po zbytek práce. Důležité je nezaměňovat jej s tzv. „technologickým forecastingem“, o kterém je následující podsekce.

1.4.1 Technologický forecasting

Technologický forecasting, anglicky „Technology forecasting“, se zaměřuje na předpovídání parametrů (charakteristik) technologií. Jedná se o parametry jako výkon procesorů, přesnost měřících zařízení nebo životnost baterií. Technologický forecasting odpovídá na otázky jako například „Jaký bude v roce 2018 výkon nejlepších grafických karet?“ nebo „Jak velkou kapacitu bude mít průměrný pevný disk na trhu v roce 2020?“

Pro technologický forecasting se používá jak kvalitativních, tak kvantitativních metod. Z kvalitativních metod jsou to například expertní odhady pomocí delfské metody nebo sestavování scénářů, z kvantitativní jsou to časové řady.

1.4.2 Forecasting nových technologií

Mezi technologickým forecastingem a forecastingem nových technologií je velký rozdíl. Jsou v jistém slova smyslu přímo opačné. Forecasting nových technologií se existujícími technologiemi příliš nezabývá. Naopak se zabývá novinkami, které ještě nejsou v produkci (na trhu) k dostání. Neklade důraz na jejich výkonnostní parametry, ale spíše na nové myšlenky, které implementují.

Není exaktně přijaté názvosloví pro rozlišování těchto dvou typů forecastingu, proto lze někde narazit na články nesoucí podobný název, ale zabývajícími se odlišnými typy forecastingu.

Práce související s forecastingem nových technologií a další informace lze nalézt pod následujícími anglickými klíči: „forecasting of new technologies“, „Tech mining“, „Technology mining“, „Forecasting emerging technologies“, „Emerging technologies“, „Innovation forecasting“ nebo „Technology monitoring“.

Definice forecastingu nových technologií

Následující definice forecastingu nových technologií se nachází v článku „Methods for Bibliometric Analysis of Research: Renewable Energy Case Study“ [7], více detailů o tomto výzkumu se nachází v části 1.4.3.2. Forecasting nových technologií je proces formování predikcí o budoucím stavu technologií.

Esence forecastingu nových technologií je hledání nově vzniklých výrobních postupů nebo průmyslových vynálezů. Nově vzniklých znamená, že nejsou známy laické veřejnosti a jsou často teprve ve vývoji, nikoliv v produkci. Takto chápe forecasting nových technologií autor této práce na základě studia článků zabývajících se touto problematikou.

Data pro forecasting nových technologií

Data pro bibliometrii poskytují databáze patentů, recenzovaných článků a knih. Poskytovatelé dat pro bibliometrii se nazývají zdroje a jsou to například Google Patents nebo Google Scholar. Tyto zdroje budou detailněji popsány v kapitole 3. V těchto zdrojích se nacházejí data, s pomocí nichž lze provádět forecasting nových technologií.

Tato práce se zabývá právě tématem volby zdrojů pro forecasting nových technologií. Následující podsekce se věnuje rešerši stavu tohoto typu forecastingu a vysvětluje motivaci pro výběr tématu.

1.4.3 Automatizovaný forecasting nových technologií

Forecasting nových či vznikajících technologií byl prováděn ve firmách experty. Velká změna přichází až v poslední době, kdy je snaha ho provádět automatizovaně, přičemž využívá bibliometrii, která je popsaná v podsekci 1.3.1.2. Úkolem této podsekce je odpovědět na otázky, jakým způsobem ho lze implementovat a jaké jsou jeho možnosti.

Automatizovaný forecasting nových technologií je aktuální záležitost, průzkum ukázal, že více než 70 % odborných studií bylo publikováno po roce 2010 [8].

1.4.3.1 Počátky

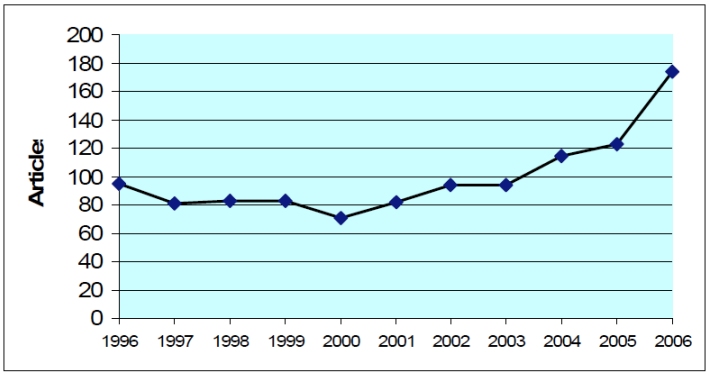
První odborné práce o „tech mining“ spojeném s bibliometrií jsou na internetu dohledatelné již z doby před rokem 2000, zejména z devadesátých let dvacátého století. Průkopníkem a zároveň tím, kdo podnítil zájem o tento obor, je Alan L.

Porter z Georgia Institute of Technology. Jeho výzkum forecastingu technologií trvá přes 25 let. Je dokonce nazýván pionýrem tohoto oboru [8].

Mezi jeho význačná díla patří „Forecasting and Management of Technology“ z roku 1991, kde diskutuje témata prognózování a řízení technologického vývoje ve firmách. Představuje metody a postupy, pomocí nichž řízení provádět. Kniha obsahuje mnoho případů z praxe a příkladů. Metody jsou tak lépe uchopitelné. V roce 2011 se svými kolegy sepsal aktualizované vydání. V knize jsou popsány postupy sběru dat od expertů, jejich vyhodnocení a následná implementace vzniklých myšlenek. V knize jsou popsány dvě možnosti získávání dat pro forecasting – sběr názorů od expertů a manuální sběr dat z databází na internetu [9]. To je velmi zajímavý způsob, protože lze automatizovat a provádět tak v mnohem větším měřítku, a to je přesně směr, jakým se forecasting nových technologií vyvíjí, jak je popsáno na následujících řádcích.

Porter také vystupuje jako jedna z vedoucích osob v „Technology Futures Analysis Methods Working Group“, která stojí za položením základů pro procesy a metody používané pro předvídání vznikajících technologií.

Pro ilustraci zájmu o predikce vznikajících technologií je přiložen následující obrázek ilustrující počet článků věnující se tomuto tématu. Následující část textu o výzkumu na MIT poskytne k obrázku zajímavou interpretaci.



Obrázek 1.1: Počet článků zabývajících se analýzou budoucích technologií [1]

1.4.3.2 Výzkum MIT a MIST

Roku 2007 byl na Massachusetts Institute of Technology (MIT) ve spolupráci s Military Institute of Science and Technology (MIST) spuštěn projekt pro vývoj analytického nástroje pro automatizovaný forecasting nových technologií. Jejich práce je zdokumentovaná ve více než deseti odborných článcích. Výzkum trval dva roky a dle přečtených článků se na něm podílelo přes dvacet lidí.

Výsledný analytický nástroj má fungovat tak, že sesbírá data, extrahuje relevantní výrazy a statistiky, vypočítá indikátory růstu a na závěr data pro lepší čitelnost rozdělí do skupin podle klíčových slov. Jsou využity principy data miningu, bibliometrie a sémantických technologií. Jednoduše nalezne výrazy, které jsou v databázích nové a opakují se v čase čím dále tím častěji. Tyto nové výrazy by měly být právě nově vzniklé technologie.

Nástroj (či systém) je rozdělen do tří logických implementačních celků:

1. sběr a agregace dat, extrakce výrazů,
2. výpočet indikátorů růstu,
3. rozdělení do skupin dle klíčových slov.

První krok bude vložení oboru zájmu do nástroje, což může být například „nanotechnologie“ nebo „obnovitelné zdroje“, vše anglicky. Nástroj potom zpracuje požadavek již nastíněnými kroky.

K pochopení principů a fungování nástroje budou všechny tři části popsány. Je to důležité zejména pro pochopení motivace vzniku a zasazení do kontextu této práce. Popis zároveň slouží k pochopení možností automatizovaného forecastingu nových technologií.

Sběr a agregace dat, extrakce výrazů

Pro sběr dat je užita metoda data miningu – bibliometrie. Jako vhodný zdroj dat pro bibliometrii byly výzkumníky z MIT a MIST identifikovány patenty a odborné recenzované články, konkrétně jejich abstrakty. Patenty jsou právní dokumenty s garantovanými právy. Jsou jim garantována práva na exkluzivitu, jež opravňuje autora k nakládání s patentem dle vlastního uvážení – lze je prodávat nebo povolit k volnému užití. Tato zákonná doba ochrany expiruje v různých právních systémech po různé době, zpravidla zhruba po dvaceti letech. Patenty mají vysoce unifikovanou strukturu, která obsahuje název, abstrakt, datum publikování, vynálezce. Mezi patenty také lze vyhledávat, což ještě více ulehčuje zmíněná struktura [10].

Odborné recenzované články si také často drží strukturu, ale nejsou tak přísně strukturované jako patenty, proto vyžadují více pozornosti při zpracování. Patenty a odborné články budou souhrnně nazývány dokumenty.

Jako první bylo třeba dle „Research plan for semantics-based technology forecasting: a case study on alternative energies“ [11] určit parametry (features), na základě kterých budou vybrány relevantní dokumenty pro zadaný obor zájmu. Tyto parametry jsou:

- Počet nalezených dokumentů
- Počet spolu-citací (co-citations)

  Jde o počet dokumentů, ve kterých jsou dvě práce citovány současně.

- Počet autorů

  Jde o počet autorů ve vzorku dokumentů, slouží k zjištění šíře zájmu o obor.

- Průměrný rok vzniku dokumentů

- Počet spoluautorů pro dokument

Výše popsané parametry jsou důležité vzhledem k práci v dalších kapitolách, kde k nim bude přihlédnuto.

Jako zdroj dokumentů sloužila databáze Scopus popsaná v podsekci 3.2.3, v krátkosti je to kvalitně udržovaná databáze citací od firmy Elsevier. Kvalitně udržovaná znamená, že do ní pravidelně přibývají dokumenty a snaží se udržovat danou strukturu, tedy přiřadit ke každé práci parametry (autoři, počet citací, abstrakt). Jako zdroj dat pro nástroj byly také zkoušeny jiné alternativy, jako Web of Science (popsaná v 3.2.4) a Google Scholar (3.2.1), ale Scopus byl autory nástroje vyhodnocen jako nejlepší možnost pro potvrzení funkce myšlenky automatizovaného forecastingu nových technologií [12]. Dále byla použita databáze Scirus, která ale po skončení tohoto výzkumu ukončila svoji činnost, čímž znemožnila fungování nástroje. Tento fakt je motivací pro vznik této práce a v další kapitole bude ještě diskutován.

Po vyhledání relevantních dokumentů nástroj provádí extrakci výrazů a sémantické sjednocení.

Extrakce výrazů (termů) je proces generování klíčových slov, na které se forecasting zaměří. Výrazy byly extrahovány z abstraktů prací, které jsou ze zdrojů skoro vždy dostupné. Extrakce byla provedena ve třech krocích, kde byly odstraněny redundantní výrazy, kterými byly generické označení vydavatele práce – jako například „journal“ nebo „conference“, dále zeměpisné názvy a na závěr byl ručně sestaven seznam informačně redundantních výrazů obsahující například „test“ nebo „manager“.

Výrazy bylo potřeba standardizovat, následovalo sémantické sjednocení – což znamená sjednocení formátů typů dat, jako jsou například data nebo převod jednotek do jedné soustavy (například libry na kilogramy).

Důležitý pro tuto závěrečnou práci je fakt, že implementace napojení na každý zdroj je časově náročná činnost, a tak ji nelze provádět metodou pokus – omyl. Motivací pro vznik této práce je pomoci budoucím zájemcům o forecasting nových technologií s výběrem vhodného zdroje pro dosažení lepších výsledků prognózování. Negativní ekonomické dopady při špatném výběru zdroje dat jsou popsány v sekci 2.3.1.

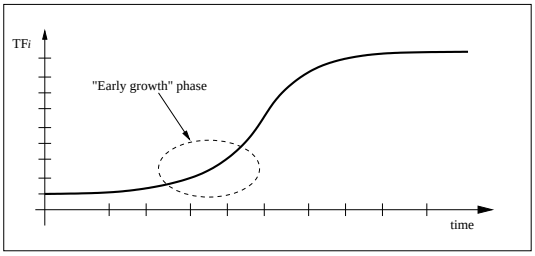
Výpočet indikátorů růstu

V této části je vypočítán počet výskytů výrazů oproti minulým letům a na základě průběžných výpočtů je nástrojem sestaven graf. Z toho je potom patrné, jestli technologie vzniká a zda je o ní zvyšující se zájem. Jako měřítko

je vybráno „Term frequency“ (TF_i), tedy počet výrazů i v dokumentu, který je definovaný jako:

$$TF_i = \frac{n_i}{\sum_{j \in I} n_j} \tag{1.5}$$

kde n_i je počet výskytů výrazů i, I je množina výrazů nacházející se ve všech zkoumaných abstraktech v jednom roce. Tento výpočet je proveden vždy pro jeden rok, když se výsledky za jednotlivé roky vynese na časovou osu, vznikne graf zájmu o technologii v čase. Ta je vyobrazena na obrázku 1.2. Na stejném obrázku je také vyznačena část, která se označuje jako „early growth“ fáze, fáze raného růstu, což značí právě fázi vznikající technologie, kterou se práce snaží u technologií identifikovat. Tato fáze raného růstu se zleva dotýká inflexního bodu grafu.



Obrázek 1.2: Fáze vznikající technologie [12]

Oproti originálnímu grafu původem z citovaného článku byly do výše zobrazeného grafu přidány popisky os – spodní osa reprezentuje čas a levá osa reprezentuje Term frequency pro zkoumaný výraz. Zajímavou paralelu lze nalézt s obrázkem 1.1 zobrazující zájem o predikci nových technologií. Lze vypozorovat, že samotná predikce nových technologií byla roku 2006 ve fázi raného růstu (early growth).

Předmětem dalšího výzkumu zůstávají další růstové indikátory, což je způsob, jak identifikovat fázi vznikající technologie. Data obsahují mnoho nedokonalostí, proto je třeba indikátory vylepšovat. V datech bývá šum nebo nejsou k dispozici data za delší časový úsek. Hlavním problémem je, že trend je třeba rozpoznat v začátcích a ne tehdy, kdy je jasně viditelný, jako na zobrazeném grafu – to už může být o několik let později.

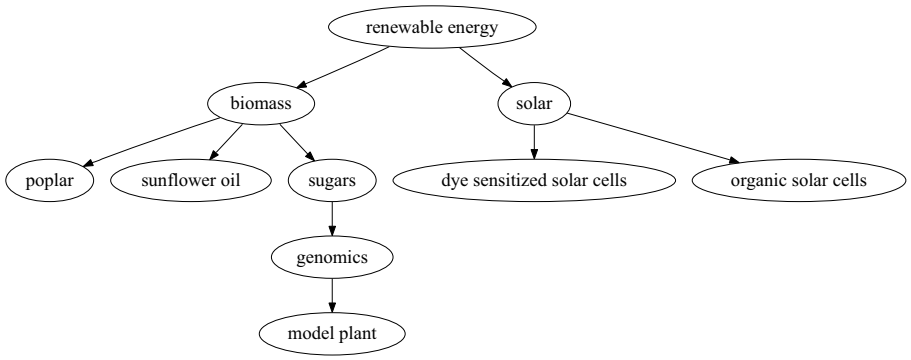
Generování taxonomie

Výsledkem předchozích operací je množina výrazů ohodnocených počtem výskytů v dokumentech v závislosti na roce. Problémem ale je identifikovat konkrétní vznikající technologii a nejen její obor nebo nadřazené pojmy, tedy eliminovat šum. Pro vysvětlení slouží následující příklad.

V oboru obnovitelných zdrojů ku příkladu vznikla technologie organických solárních článků. Dosavadní popsaný postup detekuje nárůst slov jako „obnovitelné zdroje“ (renewable energy), „sluneční energie“ (solar power), „organické solární články“ (organic solar cells), „biomasa“ (biomass), „cukry“ (sugars) a další.

Úlohou generování taxonomie je identifikovat, jaký z výrazů je vlastně nová technologie. Taxonomie je hierarchické srovnání výrazů relevantních ke konkrétní doméně například do stromové struktury. Detailní popis algoritmů pro tvorbu taxonomie se nachází v práci „Semantic Distances for Technology Landscape Visualization“ [13]. Poskytnout další popis těchto algoritmů není cílem této práce.

Výsledná taxonomie pro výrazy související s obnovitelnými zdroji vypadá následovně.



Obrázek 1.3: Taxonomie pro obnovitelné zdroje [12]

Výše jsou v hierarchii nadřazené pojmy, které zaznamenaly nárůst společně s níže položenými, cílem forecastingu nových technologií je ale odhalit právě ty níže lokalizované, tedy listy na zobrazeném stromě s výškou 4.

1.4.3.3 Další možnosti implementace

Předchozí výzkum popisoval základ a hlavní funkční prvky automatizovaného technologického forecastingu. Zmíněná struktura bývá základem pro forecastingové nástroje s tím, že se některé části pouze obměňují, čímž vznikne nástroj nový. Tedy mnoho z užitých prvků lze obměňovat a vytvářet tak nástroje s jinými vlastnostmi. Dle „The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis“ [8] existuje více možností zkoumání výrazů – podle syntaxe nebo sémantiky. Syntaxe využívá spíše statistické přístupy, kdy nezkoumá přímo význam výrazů, ale tvoří síť, kde podobné výrazy páruje a prohledává za účelem nalezení nových výrazů. Na druhou stranu

sémantický přístup zkoumá význam, tokenizuje výrazy a hledá pro ně synonyma.

Následuje shlukování, kde systém hledá shluky dokumentů navzájem podobných a zároveň různých od ostatních. Pro tyto účely jsou implementovány algoritmy jako Naivní Bayes (Naive Bayes, NB), umělé neuronové sítě (Artificial Neural Networks, ANN), logistická regrese (Logistic Regression, LR), Support Vector Machines (SVM) nebo náhodný les (Random Forest, RF) [14].

V dnešní době je výzkum predikce nových technologií stále populárnější, na popsaný základ v předchozí podsekci jsou aplikovány různé metody, kdy snahou je dostat co nejpřesnější predikce.

Kapitola 2

Ekonomické přínosy a dopady

Tato kapitola popisuje ekonomické přínosy a dopady prognózování s důrazem na automatický forecasting. Prognózování hraje v dnešní době významnou roli v podnikové sféře, zejména pro strategii podniku.

První část této kapitoly se věnuje ekonomickým dopadům prognózování pro podniky. Tato část začíná vysvětlením strategie podniku a obecných pojmů souvisejících s provozem podniku a vlivem forecastingu na ně. Dále se první část věnuje firmám, které provozují prognózování jako hlavní zdroj činnosti. V této části je kladen důraz na automatický kvalitativní forecasting, pro kompletnost je uveden jeden příklad s kvantitativním forecastingem.

Druhá část kapitoly se věnuje podnikům, které poskytují služeb forecastingu. Rozdělení typů firem v sekci 2.2.1 je výsledkem analýzy autora. Je to originální rozdělení, které se poprvé objevuje v této práci.

Třetí část této kapitoly se věnuje konkrétně automatizovanému forecastingu nových technologií, který je definován v minulé kapitole v sekci 1.4.3, a jeho dopadům při strategickém plánování budoucího směřování podniku a zavádění nových technologií.

Popis dopadů forecastingu je rozdělen na dvě logické části (v této kapitole se části nachází v první a třetí sekci) proto, že forecasting a jeho přínos pro podniky se dá chápat právě ve dvou rovinách. První rovina, které odpovídá první část, předpovídá číselné ukazatele, jako například očekávaný obrat firmy nebo počet prodaných produktů, a často vychází z matematické statistiky. Tyto výpočty obstarává automatizovaný forecasting implementací kvantitativních metod. Pro podniky je významný také neautomatizovaný forecasting, jež je výsledkem aplikace některých z kvalitativních metod forecastingu popsaných v sekci 1.3.2. Konkrétně se jedná o prognózování u řízení rizik. Druhá rovina, jejíž ekonomické dopady jsou popsány ve třetí části, je právě automatizovaný forecasting nových technologií, který se začíná v posledních letech promítat do rozhodování firem, a to zvláště díky implementaci automatizovaných nástrojů pro jeho provozování, díky čemuž je pro firmy dostupnější.

2.1 Role forecastingu ve strategii podniku

Jak je naznačeno v úvodu této práce, forecasting je předvídání budoucího stavu. Každá firma se musí v nějakou chvíli podívat do budoucnosti a rozhodnout se, jak bude postupovat – vyrobí více či méně produktů nebo nabere více či méně lidí. Předvídání je přirozená činnost, provádí ji všichni, jednotlivci, malé i velké firmy.

2.1.1 Strategie podniku

Čím jsou firmy větší, tím více se musí dívat do budoucnosti a vytvářet si plány pro budoucí chod. Intuitivně lze odvodit, že je to podmíněno potřebou finančních prostředků, velká firma potřebuje k udržení velký objem financí a pro velký objem financí je třeba dlouhodobý plán, protože velké zisky zpravidla nepřichází nahodile. Dalším důvodem pro vypracování takového plánu je určení vize, kam podnik směřuje a jak má být řízen. Takové plány se nazývají strategie podniku a podkladem pro ně jsou i prognózy budoucího stavu různých ukazatelů.

Co je to tedy podniková strategie? Je to dlouhodobý směr a rámec organizace. Strategie přináší výhodu v měnícím se prostředí díky nastavení zdrojů a kompetencí s cílem plnit očekávání zainteresovaných stran (stakeholders) [15].

V jiných zdrojích se objevují další pravdy o strategiích, například že jsou to také kritická zhodnocení podniku, analýza současného stavu a stanovení vize, v rámci které management vymezuje směřování podniku a jeho pozice v budoucnosti [16].

Strategie podniku nastavuje strategické cíle. Jsou to konkrétní body, kterých má být v období platnosti strategie dosaženo. Takovým cílem může být například otevřít novou pobočku v Evropě nebo zdvojnásobit čistý zisk.

Soulad se strategickými cíli kontroluje a ovlivňuje řízení výkonnosti.

2.1.2 Řízení výkonnosti

Řízení výkonnosti neboli Corporate Performace Management (CPM) zabezpečuje aktivity pro kontrolu a měření výkonnosti celé firmy. CPM bývá označováno jako další generace BI (Business Intelligence), což je „sada procesů, aplikací a technologií, jejichž cílem je účinně a účelně podporovat rozhodovací procesy ve firmě“ [17].

CMP zahrnuje metodiky pro efektivní řízení podniku a pro jejich implementaci. Metodiky CMP jsou například Balanced Scorecard (BSC), SWOT analýza, Activity Based Costing (ABC) nebo Theory Of Constraints (TOC). Metodik existuje mnohem více, všechny slouží pro podporu managementu a řízení podniku.

CMP dále zahrnuje metriky, které jsou definované v rámci metodik. Ty se nazývají ukazatelé výkonnosti a existují tři typy – KRI (Key Result Indicators), PI (Performance Indicators) a KPI (Key Performance Indicators). Liší se v tom, jak detailně výkonnost firmy mapují, KRI je nejobecnější, bude ku příkladu kontrolovat celkový zisk, KPI je nejdetailnější, bude sledovat mnoho parametrů, které k celkovému zisku vedou, mezi nimi například počet zaměstnanců.

Procesy nebo také komponenty CPM jsou stavební prvky, ze kterých se CPM skládá. Všechny komponenty jsou významně propojené a souvisí spolu. Mezi komponenty CPM patří komunikace, monitorování, reportování, řízení výdajů, analýzy a v neposlední řadě právě prognózování. U prognózování jde o vyhodnocování scénářů, identifikaci trendů, kontrola rozpočtů s přesahem do budoucnosti a zejména aktualizaci plánů.

CPM je se strategií podniku silně provázána, přeci jen jeden z důvodu jejího vzniku je lepší kontrola plnění strategických cílů. Řízení výkonnosti kontroluje implementaci plánů pro plnění cílů. A právě tyto plány jsou velmi významně ovlivňovány prognózami.

Plnění strategických cílů

Ke strategickému cíli vede cesta vytyčená plány. Ty jsou průběžně v čase vyhodnocovány na základě historických dat a také na základě prognóz, což je právě úloha CPM. Je vyhodnoceno, jestli plány vedou ke splnění cíle či nikoliv a dle toho jsou jsou upraveny tak, aby směřovaly k uskutečnění vytyčených cílů.

Pro lepší představu poslouží krátký příklad. Jeden z bodů podnikové strategie podniku na prodej textilu je snížit počet produktů na skladě, protože se tam často hromadí a neprodají se. Podnik prodává přes 1000 typů textilu (různé barvy, střihy atp.). Pro výpočet tohoto příkladu lze užít automatizovaný kvalitativní forecasting, konkrétně Seasonal Average approach. Stačí sesbírat data za posledních pár let, dát je do souvislosti s aktuálními daty prodejů a výsledkem je graf, který předpovídá prodeje textilu po měsících. Pro každý jednotlivý produkt bude předpověď samostatně na každý měsíc. Díky tomu je možné lépe naplánovat produkci a lépe rozdělit výrobu. Produkty se nebudou hromadit neprodané nebo nebudou v prodeji chybět a výroba bude efektivnější. Na druhou stranu konkurenční podnik, který odhaduje prodeje bez nástroje pro forecasting bude mít vyšší náklady jak za uskladnění, tak mu mohou v prodeji některé produkty chybět a může tak ztratit zákazníky. Tento velice jednoduchý příklad ukazuje modelové užití predikce ve vztahu k plnění cílů strategie podniku.

Problém s prognózováním je, že čím dále do budoucnosti se díváme, tím je to těžší. Do hry se vkládá více a více proměnných a neočekávaných jevů, které mohou výrazně zamíchat s tržním prostředím. Významným prvkem strategie je proto změna.

2.1.3 Řízení rizik

Riziko je míra nejistoty, přičemž v různých podnikových oblastech existuje mnoho druhů rizik. Analýzou a snižováním rizik se zabývá obor řízení rizik.

Řízení rizik má několik fází, jsou to identifikace, analýza, zhodnocení, ošetření, zvládnutí a monitoring rizik. Popsány budou právě první dvě fáze, tedy identifikace a analýza, které užívají prognózování.

Výstupem prvních dvou fází má být číslo neboli dopad, jaký riziko má. Výpočet je funkcí pravděpodobnosti, která vyjde z fáze zhodnocení, a z důsledku, který je výstupem analytické fáze.

Fáze, které jsou závislé na prognózování, jsou identifikace a analýza rizik. Ty jsou prováděny kvalitativními metodami, jako je například brainstorming, nebo sestavování scénářů. V takových případech se sejdou experti na danou doménu a zvažují nad alternativními možnostmi pro vývoj projektu, trhu nebo provozu podniku. Při analýze odhadují pravděpodobnost identifikovaného rizika. Výstupem prvních dvou fází je dopad.

Řízení rizik operuje s kvalitativním forecastingem, který nelze automatizovat, jelikož se opírá o interakci s experty a o jejich kreativní názory.

2.1.4 Ekonomické dopady

V této podsekci jsou shrnuty ekonomické dopady řízení výkonnosti a řízení rizik. Jak je u každého tématu naznačeno, každé je nějakým způsobem spojeno s prognózováním.

Na základě příkladu z podsekce 2.1.2 o řízení výkonnosti je jasně vidět, že díky prognózování může podnik lépe plnit své strategické cíle a plnění těchto cílů je přímo spojené s ekonomikou podniku. Příklad se omezil na přijímání zaměstnanců, ale předvídání funguje podobně i pro finanční aktivity podniku, stačí si do příkladu místo zaměstnanců dosadit počet prodaných produktů. Iniciativa pro zavedení CPM dokonce často přichází od finančního ředitele podniku [17].

Úspěšné řízení rizik má výrazné ekonomické dopady na fungování podniku. Schopnost předcházet predikovatelným rizikům je pro podniky klíčová. Proto je také řízení rizik definováno jako jeden z ISO standardů [18].

2.2 Forecasting v byznysu

Forecastingu se dle serveru VentureRadar (dostupný na adrese ventureradar.com) v dnešní době věnují stovky firem. Další možností ekonomického využití je tedy poskytovat jej jako službu. Mnohé se zaměřují jen na určitou oblast, jako je zemědělství, lesnictví nebo technologie. Existují i takové, které se snaží pokrýt forecasting obecně. V této sekci jsou popsány typy firem poskytující forecasting, konkrétní firmy poskytující služby forecastingu a na závěr je představena novinka v tomto oboru – forecasting as a service.

2.2.1 Typy firem poskytující forecasting

Existuje mnoho služeb a nástrojů, které firmy pro podporu forecastingu poskytují. Firmy a jejich nástroje lze dle výzkumu autora této práce kategorizovat do tří úrovní. Tato kategorizace je originální a objevuje se poprvé v této práci. Jednotlivé úrovně se liší v nutnosti pochopení procesů a byznysu cílové firmy, tedy té, pro kterou je prováděna predikce. Zároveň čím vyšší je úroveň, tím více finančních prostředků musí klientský podnik pro forecasting vynaložit, a tím lepší a přínosnější predikce lze očekávat.

Konkrétně automatizovaný forecasting (kvantitativní, tedy bez expertů) je prováděn na prvních dvou úrovních. Nástroje se zaměřují na predikci číselných hodnot na základě historických dat, jako jsou například zisky, obrat, počet prodaných produktů. Druhou možností je užití umělé inteligence a multiagentních systému k simulaci tržního prostředí tak, jak ho vnímá klientský podnik. K tomu slouží sofistikovanější nástroje popsané u druhé úrovně.

První úroveň

Pro správnou funkci nástrojů a služeb první úrovně není potřeba žádná znalost domény nebo byznysu cílové firmy. Většinou se jedná o automatizované výpočetní modely, do kterých operátoři, kterými jsou často sami zákazníci, vyplní historické statistiky a snaží se z nich dostat predikce. Jsou silně závislé na zkušenosti operátora a na kompletnosti statistiky. Jsou zpravidla nejlevnější, protože nevyžadují součinnost na straně poskytovatele. Pokud ano, tak jde o placenou podporu. Příkladem může být Microsoft Excel, který má implementované modely pro predikce, do kterých operátor vyplní historické statistiky a provede výpočet predikce. Většina firem se věnuje vylepšování modelů, ale myšlenku naplnění historickými daty a výpočet budoucích hodnot na základě statistických metod všechny tyto firmy sdílí.

Druhá úroveň

Jako druhá úroveň jsou klasifikovány sofistikovanější a dražší nástroje, se kterými klientský podnik pracuje. Může se jednat o tzv. Decision support systémy – systémy pro podporu rozhodování. Ty zvládají funkce první úrovně a často přidávají umělou inteligenci nebo multiagentní systémy k simulaci trhu tak, jak je relevantní pro klientský podnik. Tyto systémy slouží k podpoře klíčových rozhodnutí pro manažery podniků. Využití najdou například při uvedení nového produktu na trh, jak uvádí studie „A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: a comparative analysis“ [19]. Firmy poskytující tyto nástroje musí znát doménu klienta, ale stále lze použít obecných algoritmů k vytvoření predikcí.

Další z pokročilých nástrojů, které firmy druhé úrovně užívají jsou také nástroje pro automatizovaný forecasting nových technologií. Toto je dobrý příklad, jak se posunují hranice automatizace a strojového zpracování, jelikož

automatizovaná predikce nových technologií je mladá oblast provozovatelná firmami druhé úrovně, přičemž se vyčlenila z neautomatizovaných expertních analýz, které provozují firmy třetí úrovně.

Třetí úroveň

Na třetí úrovni už se podniku a jeho výzvám věnují analytici a experti, kteří jsou s podnikem dobře seznámeni. Firmy, zabývající se forecastingem, lze nazvat také analytické firmy nebo poradenské firmy. Do jejich práce jsou často zahrnuty předchozí dvě úrovně, aktivity třetí úrovně ale nejsou automatizované.

Snahou technologických firem zabývajících se forecastingem je posunout hranice druhé úrovně ke třetí a čím dál více aktivit třetí úrovně automatizovat. Znakem takového technologického pokroku je například vytvoření IBM Watson Health, který je představen v podsekci 1.3.3.

2.2.2 Forecasting as a Service (FaaS)

V posledních dvaceti letech nastává rozvoj modelu „.. as a service“. Infrastruktura, software nebo IT je dodáváno jako služba. Vznik tohoto fenoménu umožnily cloudové služby a dostupné připojení k internetu. Dalším krokem pro forecasting tak může být forecasting as a service. Zmínky o forecasting as a service lze dohledat pouze mladší deseti let.

Jednou z největších výzev je řešit nesourodost vstupních dat. Ta musí být předzpracována, než je z nich možno produkovat predikce. Předzpracováním je myšleno pročištění dat od chyb, výběr vhodných záznamů, převedení do proměnných a reprezentace v databázi.

V práci „Sales Forecasting as a Service“ [20] od autorů z univerzit v Nizozemsku je prezentována implementace myšlenky, jak propojit forecasting modul s informačním systémem tak, aby byl dostupný v cloudu jako služba. Problémy s předzpracováním dat řeší informační systém a ke komunikaci s forecasting modulem využívá API (Application Programming Interface). Modul předpovídá prodeje e-shopu s oblečením a dle citované práce se podařilo splnit nastavené podmínky pro to, aby se modul mohl nazývat službou (as a service).

2.2.3 Firmy poskytující služby forecastingu

Mnoho firem se soustřeďuje na hledání nových cest pro automatizovaný forecasting a s tímto námětem také vzniká mnoho firem. Pro získání představy stačí navštívit stránku VentureRadar, která se zabývá průzkumem mladých a inovativních firem, s klíčovým slovem „forecasting“ (dostupný na adrese www.ventureradar.com/keyword/Forecasting).

Velkých firem, které se etablovaly jako lídři na této pozici, je méně. V následujícím textu je uvedena jedna z nejznámějších firem – Gartner. Historicky se jednalo o konzultantské firmy, které sledovaly trendy, a pro udržení své pozice

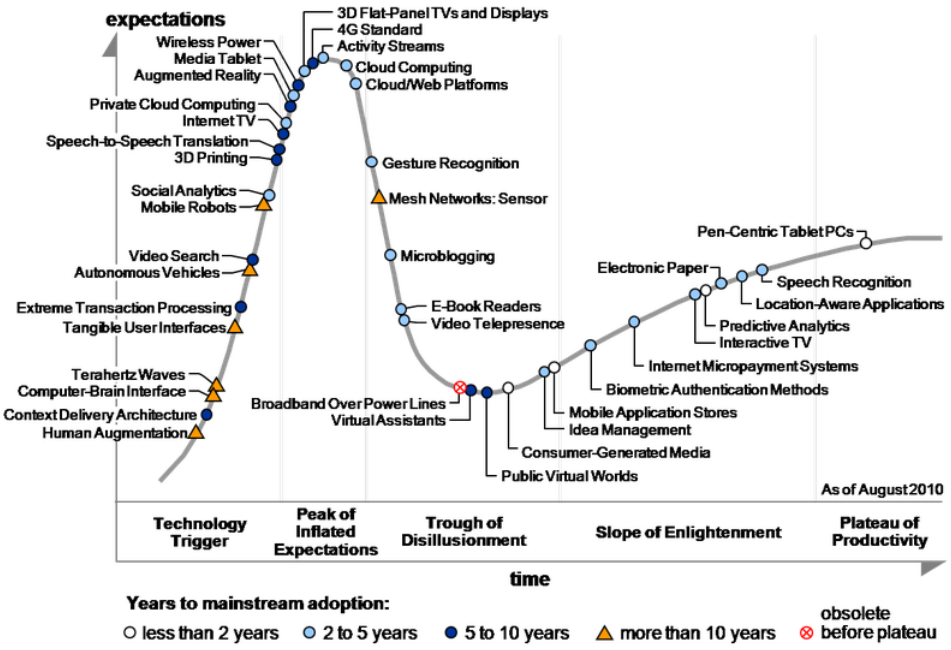
na trhu vyvinuly své nástroje pro automatizovaný forecasting. V jejich portfoliu je poskytování služeb forecastingu všech úrovní definovaných v podsekci o typech firem poskytující forecasting.

Gartner

Gartner je IT poradenská firma původně sídlící v USA založená roku 1979. Jejími zákazníky jsou ředitelé IT firem (CIO – Chief Information Officer), velké korporace, vládní agentury, technologické společnosti a velké investorské skupiny. Dle svých reportů má přes 8100 zaměstnanců v 85 zemích.

Firma je dobře známá díky svému „Hype cycle“, životnímu cyklu nových technologií. Je to graf, který je vydáván každý rok. Ten popisuje pět fází každé vznikající technologie – počátek (Technology Trigger), nejvyšší očekávání (Peak of Inflated Expectations), rozčarování (Trough of Disillusionment), postupný růst (Slope of Enlightenment) a ostré nasazení (Plateau of Productivity). Každá fáze má své části, širší popis je dostupný na oficiálním webu Gartner Hype cycle [21].

Forecasting nových technologií se snaží předpovídat technologie, které jsou na Hype cycle ještě před první částí první fáze, což je výzkum a vývoj. V následujících kapitolách bude Hype cycle hrát důležitou roli. Následující obrázky ukazují příklady Hype cycle grafů. Konkrétně grafy z roku 2010 (2.1) a z roku 2016 (2.2) budou v kapitole 4 důležité kvůli umístění autonomních vozidel (autonomous vehicles) a 4D tisku (4D Printing) na grafu.



Obrázek 2.1: Hype cycle 2010 [22]

Obrázek 2.2: Hype cycle 2016 [23]

Gartner je veřejně známá díky grafům Hype cycle, jako poradenská firma ale zejména dodává svým zákazníkům rady a nástroje, které umožňují měřit a plnit jejich nastavené cíle. Zaměřuje se na čtyři oblasti, na základě kterých dokáže tvořit analýzy – zákazníci, trhy, technologie a konkurence [24].

Firmy poskytující FaaS

Mezi firmy poskytující FaaS nástroje se řadí například VanguardSW (dostupný na adrese www.vanguardsw.com) nebo Chainalytics (dostupný na adrese www.chainalytics.com). Druhá zmíněná zavádí pojem SFaaS, což je dle její interpretace Statistical Forecasting as a service. Je to stejné, co je popsané v předchozí podsekci, jen je v názvu zdůrazněna role statistických metod. V blízké době se dá očekávat nárůst počtu takto zaměřených firem, jelikož FaaS je novinka posledních let a přesun nástrojů forecastingu do formy „as a service“ je logickým krokem ve vývoji tohoto oboru, stejně jako mnoha oborů před ním.

2.3 Forecasting nových technologií v byznysu

Analýza nových a vznikajících technologií je zejména pro technologické firmy klíčová, a to z několika důvodů. Motivací jsou samozřejmě ekonomická hlediska, což je buď ušetření finančních prostředků nebo navýšení výdělků. A nejde o ovlivnění malých částek, americké podniky utratí ročně přes 100 miliard

dolarů na výzkum a vývoj [25]. A výzkum a vývoj se vyplatí – podniky spolupracující v rámci Amerického průmyslového výzkumného institutu (Industrial Research Institute) udávají, že 35 % jejich zisků pochází z produktů vyvinutých za posledních pět let. To znamená, že z každých vydělaných tří dolarů je jeden získaný díky nasazení aktuálních technologií.

Přehled o vznikajících technologiích je důležitým zdrojem informací pro inovátory, výzkumné inženýry, ředitele podniků i tvůrce legislativ. Mají význam na nadnárodní úrovni (například Evropská unie) až po individuální organizace. Na následujících řádcích je několik konkrétních příkladů, kdy se vyplatí znalost vznikajících technologií.

Vědecký výzkum

Výzkumníci se snaží přicházet s novými technologiemi a vylepšovat stávající. Jejich práce musí být co nejefektivnější. Potřebují vědět, jestli se někdo ve světě nevěnuje stejnému tématu jako oni – existuje pravděpodobnost, že by si mohli vypomoci, vyměnit si know-how a poradit si. V tom horším případě by už někdo mohl dokonce podobnou technologii vymyslet před nimi. Správné odpovědi na tyto otázky ušetří mnoho času a finančních prostředků.

Aplikovaný výzkum a vývoj

Jde o oblast, kde na výzkum navazuje implementace myšlenek a studie jejich uvedení na trh. Podniky potřebují vědět, jestli jsou jejich technologie komerčně využitelné a zdali se vyplatí je vůbec dotáhnout do produkce. To mohou zjistit rozhlédnutím se po podobných nových technologiích a analýzou jejich úspěšnosti na trhu.

Tato analýza pomáhá také s prioritizováním úkolů, jaké technologie ve vývoji podpořit a které naopak například přestat vyvíjet.

Podniková strategie

Podniky hledají produkty, se kterými vstoupit na trh. Využijí prohledávání vznikajících technologií pro zjištění, jaké se hodí do zaměření jejich byznysu. Díky tomu mohou naplánovat vývoj nového produktu.

Na nadcházející trendy se mohou společnosti připravit s velkým předstihem, plánování na několik let dopředu je pro podnikovou strategii charakteristické.

Podnik může narazit i na konkurenci, která se zajímá o stejnou vznikající technologii. Pokud by podniky chtěly spolupracovat, přichází na řadu několik řešení, mezi nimiž je licencování technologie druhému podniku nebo tzv. joint venture. To je forma spolupráce dvou entit, zpravidla podniků, které se spojí za účelem společného projektu, přičemž každá z firem do nového podniku přináší určitou přidanou hodnotu.

Tvorba legislativy

Tvůrci legislativy využijí znalost o nových patentech pro hledání právních děr v zákonech zemí a díky tomu mohou přijít s novelami ještě předtím, než tyto mezery objeví někdo jiný, kdo by jich mohl zneužít. Mohou tak reagovat dopředu, nikoliv retroaktivně.

2.3.1 Přínosy automatizovaného forecastingu nových technologií

Nejčastěji se forecasting nových technologií provádí kvalitativními metodami, tedy brainstormingem, tvorbou scénářů nebo delfskou metodou. Podstatou všech těchto metod je přítomnost expertů. Oproti tomu stojí automatizovaná verze, která je již popsána v sekci 1.4. Ta přináší oproti neautomatizované verzi množství výhod.

Je finančně a časově náročnější implementovat systém pro automatizovaný forecasting než zorganizovat sezení expertů. Na druhou stranu se jedná spíše o jednorázovou investici, která by neměla vyžadovat příliš výdajů do oprav. Systém vyžaduje opravy zejména v případě, že některý ze zdrojů změní model poskytování dat. Potom je nutné systém přeprogramovat, pokud by to nebylo možné a zdroj by přestal data poskytovat úplně, bylo by nutné identifikovat jiný zdroj, na který lze systém napojit. Přesně toto se stalo výzkumníkům z MIT, jejichž práce je popsána v sekci 1.4.3.2, když použili databázi Scirus, jejíž běh byl v roce 2014 ukončen. Tato závěrečná práce chce pomocí měření kvality zdrojů dat najít nejlepší z nich pro použití v automatizovaném forecastingu, a tak předejít mnohdy zbytečným výdajům na údržbu zmíněných systémů a zároveň zlepšit výsledky systému tak, že doporučí nejkvalitnější zdroje dat. Způsob provedení je popsaný v dalších kapitolách.

Systém pro automatizovaný forecasting bude generovat závěr a doporučení rychleji než sezení expertů. To je nutné svolat, provést a vyhodnotit, naopak systém dokáže generovat výsledky nikoliv v řádu dnů, ale minut. Příklad takového systému je popsán znovu v části výzkumu MIT v sekci 1.4.3.2. Díky tomu mohou manažeři provádět klíčová rozhodnutí mnohem rychleji, nemusejí čekat týdny na informace z expertního sezení.

Další výhodou je aktualita informací, oslovení experti nemohou mít informace o všech novinkách ze všech koutů světa, kdežto v databázích dokumentů (patentů, odborných prací..) se nacházejí nejnovější informace. Díky tomu lze odhalit vznikající technologii velice brzy, například ve stupni, kdy se jí věnuje pouze jedna výzkumná jednotka (firma, univerzita).

Díky exaktním číslům, jako je například počet dokumentů, ve kterém se vyskytuje klíčový výraz, nebo počet citací, lze také sestavovat grafy a přehledy, ze kterých je patrná změna zájmu o konkrétní technologii, ať už se jedná o zvýšení zájmu či snížení. To zjednodušuje interpretaci výsledků a podporuje rozhodovací procesy spojené s forecastingem.

2.3.2 Shrnutí ekonomických dopadů forecastingu

V této kapitole bylo popsáno, proč je pro firmy výhodné užívat forecasting. Implementovat si vlastní nástroje je pro naprostou většinu firem nesmyslné, jelikož pro ilustraci jen implementace nástroje z podsekce 1.4.3.2 zabrala přes dva roky přibližně deseti až dvaceti lidem. Jednoznačným řešením je tak outsourcing této aktivity, přičemž outsourcing a služby „.. as a service“ jsou v dnešní době trendem.

Firmy poskytující forecasting jsou v této kapitole v podsekci 2.2.1 rozděleny do třech úrovní. V průměrném případě by měly velikosti firem poptávající služby forecastingu odrážet zvolenou úroveň firmy poskytující forecasting. Tedy pro menší firmy budou stačit služby firmy poskytující první úroveň a naopak pro největší firmy třetí úroveň. S přibývající úrovní totiž rostou náklady, na druhou stranu ale také roste přínos. Je nutné zvážit, jestli například malá firma vůbec potřebuje samotného analytika či jí stačí jednodušší automatizovaná predikce nákladů a výnosů.

Kapitola 3

Analýza zdrojů

V předchozích kapitolách byly popsány možnosti a ekonomické dopady forecastingu obecně, forecastingu nových technologií a také jejich automatizovaných variant. Práce se bude nadále věnovat automatizovanému forecastingu nových technologií.

Autoři systémů pro automatizovaný forecasting nových technologií nevěnují příliš velkou pozornost výběru datových zdrojů, nad kterými implementují své systémy (popsané v sekci 1.4), což má výrazné ekonomické dopady (popsané v podsekci 2.3.1).

Tato práce proto prozkoumá potenciální zdroje dat a navrhne obecnou metodiku pro vyhodnocení vhodnosti jejich použití v systémech pro automatizovaný forecasting nových technologií. V této kapitole jsou identifikovány a popsány zdroje, které se dají využít pro zmíněný účel.

Datový zdroj

Datový zdroj není obecně jednoznačně specifikovaný termín. Pro potřeby této práce bude definován jako databáze často s webovým rozhraním, ve které jsou uloženy bibliografické dokumenty či bibliografická data. Bibliografickými dokumenty resp. daty jsou myšleny patenty, odborné články, závěrečné práce, knihy, reporty z konferencí, ale i například novinové články.

Tato práce rozlišuje tři skupiny datových zdrojů – zdroje s patentovými dokumenty, s odbornými znalostmi a zprávy. Jsou takto rozděleny z důvodu jejich různorodosti. Patentová data jsou sice nejaktuálnější, ale také se mezi nimi může nacházet mnoho zavádějících prací a konceptů. Na druhou stranu do novinových článků a zpráv se nedostávají ty nejnovější technologie, ale mohou se tam dostat ty, které v patentech nejsou k nalezení. Nejvýznamnější vynálezy firem totiž nebývají patentované, a to z důvodu konkurenčního boje [26]. Rozumnou střední cestou mohou být vědecké články, knihy a konferenční příspěvky.

3.1 Patentové zdroje

Definice patentového dokumentu (patentu) je uvedena v podsekci 1.4.3.2. Pro krátké připomenutí – patenty jsou právní dokumenty s garantovanými právy a mají vysoce unifikovanou strukturu, která obsahuje název, abstrakt, datum publikování, vynálezce. Zdroj s patentovými daty jsou pro bibliometrii nejvhodnější, jsou jednoduše automaticky zpracovatelné, objevují se v nich nejčerstvější informace. Očekávaným problémem ale je obtížná rozlišitelnost relevantních náznaků od velkého množství zavádějících indicií (v první kapitole zmiňovaného šumu). Tímto problémem se ale zabývají až další fáze procesu automatizovaného forecastingu, ty jsou popsané v podsekci 1.4.3.2. V Česku se o správu patentů stará Úřad průmyslového vlastnictví.

3.1.1 Espacenet

Espacenet je databáze vlastněná EPO – Evropským patentovým úřadem. Jeho celosvětová databáze obsahuje přes 100 mil. patentových dokumentů, což je ze všech zmíněných zdrojů patentů nejvíce.

Espacenet nabízí přístup k informacím o vynálezech a patentech datovaných od 19. století dodnes. Součástí jsou i české patenty.

Dostupné na: worldwide.espacenet.com

3.1.2 Patentscope

Patentscope databáze obsahuje 64 milionů patentových dokumentů včetně 3,1 milionů zveřejněných patentových přihlášek. Patentová přihláška je žádost o uznání patentu. To znamená, že v databázi jsou i patenty v procesu schvalování. Patentscope je vlastněn WIPO – World Intellectual Property Organization, což je organizace sdružující skoro všechny země světa (189 zemí).

Dostupné na: patentscope.wipo.int/search/en/search.jsf

3.1.3 Google Patents

Google Patents indexuje patenty jak od již zmíněných EPO (Evropský patentový úřad) a WIPO (World Intellectual Property Organization), tak i od dalších patentových organizací. Mezi ně patří asijské úřady (Čína, Japonsko, Korea), jejichž patenty byly přeloženy do angličtiny a jsou tak lépe vyhledatelné.

Dle oficiálních informací dokáže vyhledávat mezi více než 87 miliony patenty. Vyvstává otázka, proč je toto číslo nižší než u Espacenet. Je to proto, že Espacenet vyhledává ve většině evropských patentů, Google Patents ale pouze v patentech nejproduktivnějších zemí Evropy, co do počtu patentů.

Služba Google Patents byla spuštěna v prosinci roku 2006 a od té doby přibývá počet úřadů, jejichž patenty indexuje.

Dostupné na: patents.google.com

3.1.4 Další zdroje patentů

Free patents online

Free patents online obsahuje především americké patenty, jejich databáze obsahuje přes 11 milionů záznamů.

Dostupné na: freepatentsonline.com

Derwent World Patents Index

Jedna z největších databází je poskytována Clarivate Analytics, který se vyčlenil od Thomson Reuters. Obsahuje 75 milionů patentových dokumentů a je profesionálně spravovaná, její administrátoři se starají o správnost a aktualitu záznamů. Nabízí stahování dat přes API v předem definovaném formátu nebo vylepšené vyhledávání. Tyto služby jsou ale zpoplatněné.

Dostupné na: clarivate.com/products/derwent-world-patents-index

3.2 Zdroje s odbornými znalostmi

Zdroje s odbornými znalostmi zahrnují bakalářské, diplomové a dizertační práce, odborné práce, články, publikace, konferenční příspěvky. Ve všech zmíněných zdrojích se mohou nacházet zmínky nebo celé práce o trendech, které budou v budoucnu významné.

Tyto dokumenty jsou často recenzované. Recenzované dokumenty prošly hodnocením a kontrolou od odborníků na konkrétní doménu. I přesto ale nelze automaticky počítat s tím, že všechny obsažené informace budou bezchybné. Tato myšlenka je rozpracována v sekci 3.4.

Co se týče zpracovatelnosti, nemají vždy jednotný rámec, tudíž může být automatizované zpracování ztíženo. Je možné, že pro plnohodnotné zpracování bude třeba uplatnit metody předzpracování dat, které jsou schopny upravit vstupní data do uniformního formátu.

3.2.1 Google Scholar

Google Scholar funguje podobně jako Google Patents na principu indexování dokumentů z jiných zdrojů. Indexuje dostupné odborné články, knihy, konferenční příspěvky, bakalářské, diplomové a disertační práce, technické zprávy, a další odbornou literaturu z různých zdrojů. Řadí se tak na první příčku v počtu obsažených dokumentů v této kategorii.

Google nezveřejňuje velikost obsahu, ale odborné odhady z roku 2014 hovoří o 160 miliónech záznamů [27].

Dostupné na: scholar.google.cz

3.2.2 Vyhledávač Summon ČVUT

Vyhledávač má k dispozici 84 milionů výsledků, z toho 37 milionů článků v odborných časopisech, 28 milionů novinových článků a přes 230 000 knih. Je provozován Ústřední knihovnou ČVUT. Zajímavostí je, že obsahuje 30 milionů recenzovaných dokumentů, zbylé jsou například i novinové články, lze jej tady zařadit i do kategorie zdroje zpráv. Tam ale při pozdějším zkoumání řazen nebude, přičemž toto rozhodnutí bude vysvětleno. Summon je krabicové řešení vyhledávače (lze jej přeprodávat více zákazníkům), přičemž ČVUT si jej zakoupilo. Tento fakt lze také vypozorovat z webové adresy Summon ČVUT zmiňující poskytovatele – Serials Solutions.

Dostupné na: cvut.summon.serialssolutions.com

3.2.3 Scopus

Databáze recenzované literatury vlastněná společností Elsevier, jež funguje jako vydavatelství i správce a archivátor dokumentů. Scopus obsahuje abstrakty a citace z vědeckých žurnálů, knih a konferencí, a patenty.

Dostupné na: elsevier.com/solutions/scopus

3.2.4 Web of Science

Web of science je služba provozovaná firmou Thomson Reuters, jež vznikla v roce 2008 spojením firem Thomson a Reuters, do té doby nejvýznamnějších zpravodajských agentur. Tato firma jiné přeprodává nejnovější zprávy do celého světa. Zdroj je přístupný po přihlášení přes partnerskou instituci (i ČVUT) nebo po registraci. Obsahuje 90 miliónů záznamů – obsahově stejných jako poskytuje Scopus.

Dostupné na: apps.webofknowledge.com

3.2.5 Další zdroje textů

Google Books

Teto projekt není přímo podobný již zmíněným Google Scholar a Google Patents, jež sbírají data a vyhledávají mezi nimi. Oproti nim se Google Books snaží o vytvoření databáze naskenovaných knih, především pro účely uchování. Tyto knihy jsou buď celé přístupné nebo jen jejich úryvky. Většinou se jedná o starší knihy s již neaktivními autorskými právy, pokud jsou aktivní, lze si přístup k nim dokoupit. Knihy je možné také užít k automatizovanému zpracování, jelikož skoro každá kniha je opatřena náhledem a shrnutím obsahu srovnatelným s abstrakty článků.

Dostupné na: books.google.cz

CiteSeer

CiteSeer je oproti konkurentům o něco menší databáze, která jak indexuje články z internetu, tak přijímá originální články. Obsahuje přes 8 milionů článků a soustřeďuje se zejména IT obor.

Dostupné na: citeseerx.ist.psu.edu

Engineering Village

Engineering Village je zdroj zahrnující 12 databází a je přístupný pouze po autorizaci, a to buď vybraným příslušníkům univerzit nebo po yaplacení poplatku. Nabízí počítačovou literaturu a patenty, zahrnuje například americké i evropské patentové databáze. Nepodařilo se získat informaci o počtu dokumentů, ten je odvozen od počtu dokumentů v indexovaných databázích.

Dostupné na: engineeringvillage.com

3.3 Zprávy

I v mnohých neodborných (nerecenzovaných) článcích se objevují zmínky o nových technologiích a trendech. Zejména významné jsou tyto zprávy v souvislosti s faktem, že nejvýznamnější vynálezy nebývají patentované, jak je zmíněno v úvodu této kapitoly. V předchozí sekci je zmíněn vyhledávač Summon, který umí vyhledávat také mezi neakademickými články, zastane tedy i práci zdrojů z kategorie zpráv. Významnými agregátory novinových článků jsou Google News a Bing News.

3.3.1 Google News

Google News prohledává přes 350 milionů článků a agreguje přes 4500 světových zdrojů zpráv od časopisů, přes noviny, až po zpravodajství.

Dostupné na: news.google.cz

3.3.2 Bing News

Konkurence Google News s mnohem menší základnou – zhruba 30 milionů článků. Původně známé také jako MSN News vlastněná Microsoftem.

Dostupné na: bing.com/news

3.3.3 Reuters

Reuters byla zpravodajská agentura, která přeprodávala nejnovější zprávy do celého světa, a zanikla v roce 2008 spojením s firmou Thomson. O Thomson Reuters je zmínka již v předchozím textu v podsekci 3.2.4. Reuters nyní funguje pod změněnou firmou a její databáze obsahuje miliony publicistických článků z různých domén včetně technologické.

Dostupné na: reuters.com/search/news

3.3.4 Další zdroje zpráv

Technický týdeník

Český zástupce v přehledu zpráv s úctyhodnými šestnácti a půl tisíci technickými články může být zajímavou alternativou k zahraničním zdrojům.

Dostupné na: technickytydenik.cz/vyhledavani.php

3.4 Pochybnosti o úrovni dokumentů

Všechny dokumenty samozřejmě nejsou stoprocentně spolehlivé. Toto tvrzení naznačuje, že informace ve zdrojích obsažené nemusí být pravdivé, mohou být zavádějící či dokonce lživé. Tato sekce se snaží relativizovat význam slov jako „akademický“ nebo „odborný“ – na následujících příkladech je vysvětleno, proč dokonce ani recenzovaná literatura nemusí být kvalitní.

Důvodem zájmu o kvalitu dokumentů je výsledná kvalita predikcí, která může být nekvalitními dokumenty ovlivněna. Toto ovlivnění lze zmírnit dvěma možnostmi. První možností je myslet na tento fakt již při návrhu měření kvality zdrojů, čímž se zabývá následující kapitola. Snažit se precizně verifikovat dokumenty ve zdrojích obsažené, ze kterých predikce vycházejí, tedy kontrolovat počty jejich citací a celkovou kvalitu dokumentů. Druhá možnost je popsána již v první kapitole v podsekci 1.4.3.2 s výzkumem na MIT, kde se zkoumaný nástroj snažil zbavit informačního šumu, tedy mimo jiné i zavádějících informací z dokumentů.

Tato sekce poukazuje na fakt, že dokumenty ve zdrojích nemusí být vždy kvalitní, i když se tak tváří. Proto je nutné přijít s postupy, jak méně kvalitní dokumenty potlačit. Díky tomu budou vznikat kvalitnější predikce.

3.4.1 SCIgen

SCIgen je software vyvinutý výzkumníky na MIT, přístupný z webového prohlížeče (pdos.csail.mit.edu/archive/scigen). Stačí vložit jména údajných autorů a během chvíle SCIgen vygeneruje článek. Na první pohled vypadá vše v pořádku – obsahuje nadpisy, kapitoly, výsledky měření a grafy. Při přečtení jakékoliv části ovšem text nedává vůbec smysl.

Zajímavostí je, že Google Scholar je náchylný na toto podvodné jednání. Proběhlo několik experimentů, z nichž nejvýznamnější roku 2010, kdy bylo navzájem citováno velké množství zdánlivě odborných článků vygenerovaných SCIgenem. Tyto články se pak umístily ve významnosti před články od takového jména jako Albert Einstein. Na toto téma vznikla i studie „Academic Search Engine Spam and Google Scholard Resilience Against it“ [28].

Mezi další úspěchy uživatelů SCIgenu patří akceptování článku v roce 2008 na konferenci v Číně od fiktivního Herberta Schlangemanna, kterého stvořilo pár studentů. Tento článek dokonce dostal pozitivní recenzi, případem se zabývá nerecenzovaný příspěvek „Another fake paper has been accepted

for publication and oral presentation“ [29] na blogu vedeném pod jménem neexistujícího autora.

3.4.2 Beallův seznam

Jeffrey Beall je americký knihovník, známý především díky svému „Beall's list of predatory publishers“ – seznamu dravých či agresivních vydavatelů.

Vydavatelé vědeckých článků se na tomto seznamu mohli ocitnout, pokud naplnili některé z charakteristik poškozujících věrohodnost publikovaných prací, jako například akceptování článků bez recenzí a kontrol kvality, akceptování nesmyslných (např. SCIgen) nebo falešných článků, falšování citací, jmenování fiktivních akademiků do funkcí ve vydavatelství, agresivní přesvědčování autorů ke tvorbě dalších článků a mnoho dalších.

Beallův seznam potenciálních a pravděpodobných dravých vydavatelů fungoval od roku 2008 devět let, až do roku 2017, kdy byl smazán. Zahrnoval něco málo přes tisíc vydavatelů. Ale známá pravda je, že co již jednou bylo na internetu, bude tam již vždy. Proto jsou kopie Bellova seznamu i dnes stále dohledatelné.

3.5 Shrnutí a komparace zdrojů

V této kapitole bylo popsáno šestnáct zdrojů ve třech kategoriích. Pro získání lepšího přehledu je nabídnuta následující tabulka 3.1 s rokem vzniku a orientačním počtem dokumentů každého zdroje.

Název zdroje	**Založeno**	**Obsah**
Espacenet	1998	100 mil.
Patentscope	2005	67 mil.
Google Patents	2006	87 mil.
Free Patents Online	2004	55 mil.
Derwent World Patents Index	1995 (odhad)	75 mil.
Google Scholar	2004	160 mil.
Summon ČVUT	2013	84 mil.
Scopus	1995	55 mil.
Web of Science	2005	90 mil.
Google Books	2004	90 mil.
CiteSeer	2007	6 mil.
Engineering Village	1995	30+ mil.
Google News	2002	350+ mil.
Bing News	2009	30 mil.
Reuters	2008	–
Technický týdeník	2007	18 tis.

Tabulka 3.1: Shrnutí roku založení a obsahu databáze zdrojů

Kapitola 4

Návrh měření kvality zdrojů

V předchozí kapitole je identifikováno celkem 16 zdrojů, o jejich použitelnosti pro automatizovaný forecasting nových technologií je ale nutné ještě rozhodnout. V této kapitole bude řešen návrh, jak měřit kvalitu zdrojů, a to pro použití v automatizovaném forecastingu nových technologií.

Tato kapitola definuje metriky pro měření kvality zdrojů a odpovídá na otázku, jak se kvalita zdroje hodnotí a jak je měřitelná. Nejprve je třeba přesně definovat metodiku, dle kterých bude posuzována. Tato kapitola se tedy věnuje tématu kvality dat a potažmo zdrojů pro automatizovaný forecasting nových technologií a klade si za cíl definovat metriky pro její měření. Žádný standard pro měření této kvality dosud není definovaný zejména z toho důvodu, že tento obor zažil rozvoj až v posledním desetiletí (jak je popsáno v kapitole 1). Mít správně zvolený datový zdroj je ale z mnoha důvodů kritická část procesu predikce nových technologií (jak je popsáno v kapitole 1 a 2).

První sekce nabízí úvod do tématu kvality dat, které bude pro potřeby této práce chápáno jako synonymum ke kvalitě zdrojů. Druhá sekce provádí analýzu měření kvality dat v několika oborech. Ve třetí sekci je zavedena metodologie výpočtu samotných metrik.

4.1 Kvalita dat

Pro fungování firem je kritická část úspěchu správné nakládání s daty. Od uchovávání dat, přes jejich zpracování až po jejich analyzování. Proto také, zejména v posledních dvou dekádách, roste zájem o měření kvality dat. Tento zájem se objevuje v několika oblastech, ať už obecně pro data jako taková, big data nebo například linked data, přičemž tyto pojmy budou vysvětleny na následujících stránkách. Analytici z konkrétních oborů si stanovují vlastní metriky a ty se často prolínají. Jak již bylo zmíněno, pro automatizovaný forecasting nových technologií nejsou žádné metriky stanovené. Inspirací při jejich definování budou právě metriky a parametry dalších oborů zajímajících se o data.

Definice kvality dat

Data jsou kvalitní, když při použití uspokojí požadavky, které jsou na ně kladeny. Jinými slovy, kvalita dat záleží na zamýšleném užití stejně jako na samotných datech. K uspokojení zamýšleného užití musí být data přesná (accuracy), ukotvená v čase (timeliness), relevantní (relevant), kompletní (complete), pochopitelná (understood), důvěryhodná (trusted) [30].

Takto definuje kvalitu dat kniha studující obecně data vhodná pro užití ve firmách. Je to obecná definice a konkrétní vyjmenované parametry budou rozpracovány na následujících stránkách.

Požadavky na zdroje

Následující požadavky jsou kladeny na zdroje, potažmo na data, či dokumenty, které spravuje proto, aby mohl sloužit jako zdroj pro automatizovaný forecasting nových technologií. Na základě těchto požadavků budou definovány metriky pro jejich měření.

V předchozích kapitolách bylo popsáno, že častým a ověřeným zdrojem informací jsou expertní odhady. Zdroj by tedy měl obsahovat dostatek odborných informací (dat, dokumentů) srovnatelných se znalostmi expertů nebo je i převyšovat.

Přístupnost zdroje musí být dlouhotrvající, to znamená, že musí být zálohovaný a mít historii. Díky tomu lze předpokládat, že bude dostupný i v budoucnu a dá se s ním dále pracovat.

Pro automatizovaný forecasting je nutné, aby šel výstup zdroje zpracovávat výpočetně-technickými prostředky.

4.2 Analýza měření kvality dat

Na data jsou kladeny požadavky, které musí být měřitelné. Jak je naznačeno v úvodu, jedná se o téma, o které se již analytici zajímají. Často jsou ale uzavřeni do svých oborů nebo zaměření na jiný typ dat, než který je pro automatizovaný forecasting nových technologií potřeba. Jak tedy měřit kvalitu dat, které nám poskytují zdroje pro forecasting? Pro inspiraci poslouží metriky z jiných oborů. Ty budou potom prozkoumány z pohledu zmíněného forecastingu.

V následujících podsekcích jsou představeny jednotlivé obory a příslušné metriky pro měření kvality dat, ze kterých se lze inspirovat. Tyto obory již mají své metriky pro kvalitu dat na rozdíl od zkoumaného forecastingu definované.

Anglicky se tato problematika dá vyhledat pod klíči „data quality assurance“, „information quality assessment“ nebo například „evaluation of data sources“. Existuje související manažerské odvětví, které se nazývá řízení kvality dat – „Data Quality Management“.

4.2.1 Big data

Jako big data jsou klasifikována data, která splňují takzvané 3V. Jsou to tři metriky definované již před více než deseti lety a v dnešní době jsou doplňovány dalšími, kdy je zmiňováno pět i šest „V“ [31]. K nalezení je až sedm a dokonce i deset. 3V definovala firma Gartner ve svém článku „Beyond the hype: Big data concepts, methods and analytics“ [32]. Gartner je firma, která se shodou okolností zabývá forecastingem a je zmíněna v podsekci 2.2.3. Na následujících řádcích bude představeno prvních šest „V“, je čerpáno z citovaných zdrojů.

- **Velikost, objem (Volume)**

  Standardně se v dokumentech s tématem big data hovoří až o petabytech dat (1 PB = 1000 TB). Velikost je, jak název samozřejmě napovídá, jedna z hlavních domén big data. Stejně tak u dat pro automatizovaný forecasting nových technologií je objem důležitý. Lze odhadnout, že čím více článků se bude věnovat konkrétnímu tématu, tím bude implementace tématu blíže a blíže uvedení do praxe. Je to tedy důležitý indikátor růstu zájmu o technologií.

- **Pohyb (Velocity)**

  Data jsou tvořena v reálném čase v souladu s aktivitou uživatelů konkrétní služby, obchodu nebo například instituce. Zpracovatelé se snaží tato data analyzovat, těch je ale obrovské množství. Důsledkem čehož je zpracovat nestíhají a data rychle stárnou. Výsledky analýzy potom bývají zkreslené nebo nepoužitelné. Kromě masivních investic do infrastruktury se hledají další řešení, které by problém zpracování big data v reálném čase vyřešily. U forecastingu je toto pravděpodobně nepříliš použitelná veličina, dokumenty bývají statické a nepřibývají nezpracovatelnou rychlostí.

- **Formát (Variety)**

  Tato charakteristika zavedená firmou IBM reprezentuje různorodost a těžkou uchopitelnost big dat. Big data se nativně nenacházejí v jednoduše a uniformě zpracovatelném stavu. Nacházejí se v mnoha různých formátech, jakými jsou kromě dokumentů například polohové souřadnice, video, obrázky, data z webových prohlížečů. Vždy je potřeba data upravit do vhodného formátu a následně interpretovat. Ve vztahu k tématu práce se může jednat o zkoumání formátu dat ze zdrojů a klást si otázky, zdali je k dispozici API nebo jestli jsou data strukturovaná.

- **Hodnota (Value)**

  Pod pojmem „Value“, který byl představen firmou Oracle, se u big data nachází analýza reálného přínosu dat. Snaží se odpovědět na otázku,

jestli jsou výsledky použitelné pro byznys. Hlavním přínosem zpracování big data je pro firmy zvýšení konkurenceschopnosti a v konečném důsledku zvýšení zisku. Pokud zisk nepřichází, analýza big data neplní svou roli a je třeba ji přehodnotit.

- **Pravdivost (Veracity)**

  Pravdivost vyjadřuje důležitou veličinu – důvěryhodnost a jak se dá výsledkům vzešlých ze zkoumání big data důvěřovat. Tématu pravdivosti již byl věnován prostor v kapitole 2 a byla akcentována potřeba ji měřit. Ve vztahu k automatizovanému forecastingu nových technologií se pravdivosti blíže věnuje další část práce v podsekci 4.2.3 o linked datech.

- **Proměnlivost zdrojů (Variability and complexity)**

  Firma SAS následovala příkladu Gartner, Oracle a IBM a přišla také se svou charakteristikou pro big data – proměnlivostí. Od formátu (variety) se významově liší v tom smyslu, že proměnlivost je způsobována změnami v čase. Data přicházejí v periodických opakováních z různých zdrojů, což není predikovatelné a způsobuje to problémy se zpracováním. Tato charakteristika se pro automatizovaný forecasting nových technologií příliš nehodí, protože zdroje jsou statické a ani jejich struktura často nemění.

Existují ale i výzkumníci zabývající se kvalitou big data, kteří nemusí stoprocentně souhlasit s nastaveným standardem, myšlenky jedněch z nich jsou popsány v článku „The Challenges of Data Quality and Data Quality Assessment in the Big Data Era“ [33]. Akcentují potřebu definovat kvalitu dat pro konkrétní byznys, pro který jsou data zrovna sbírána, a ne obecně. Pro tuto práci je ale zajímavější, že rozdělují metriky do dvou vrstev, přičemž *kombinovaná kvalita* druhých vrstev určuje kvalitu metriky v první vrstvě. Tato myšlenka bude v další sekci této kapitoly použita.

4.2.2 Sekundární data

Pojem primární a sekundární data se obyčejně používá pro data figurující v průzkumech. Primární data vznikají přímým zásahem sběratele dat, slouží k vyřešení problému, otázky [34]. Sekundární data jsou další přídavné zdroje, které mohou vzniknout dedukcí z primárních dat, mohou být sesbírány pro jiné účely než řešení zadaného problému. Dají se popsat jako podpůrná.

Nizozemští výzkumníci identifikují 49 faktorů, které ovlivňují kvalitu sekundárních dat [35]. Zmiňované faktory (nebo také metriky) jsou identifikovány ve vztahu k dotazníkovým šetřením, ale lze je po mírné úpravě vztáhnout i na zdroje pro automatizovaný forecasting nových technologií a obohatit jimi seznam metrik.

Objevují se metriky, které by byly s již zmíněnými duplicitní, nesou pouze jiný název, jako například „correctness“ nebo „reliability“, přičemž obě jsou významově ekvivalentní již definované pravdivosti. Objevují se ale nové pojmy, které lze velice dobře užít i pro zmiňovaný forecasting. Těmi jsou:

- **Dostupnost, stabilita (Availability)**

  Dostupnost je doslova definovaná jako stav, kdy je systém schopen provozu. Tento požadavek lze velmi dobře vztáhnout i na provoz forecastingu napojeného na datové zdroje, zde by se jednalo čistě o technickou dostupnost a stabilitu, tedy stabilitu serverů poskytovatele dat.

- **Trvanlivost, kontinuita (Continuity)**

  Předmětem zkoumání trvanlivosti pro sekundární data je očekávaná doba, po kterou bude zdroj dat existovat. Toto je velice důležitá metrika, jejíž neprozkoumání či špatný odhad může způsobit silné ekonomické dopady. Před těmi je varováno v podsekci 2.3.1 a jsou rozebírány na několika místech této práce.

- **Ochota spolupracovat (Willingness to cooperate)**

  Ochota spolupracovat je definována jako stupeň ochoty poskytovatele dat ke spolupráci. Tento parametr se skvěle hodí i pro zmiňovaný forecasting. Předmětem domluvy mezi provozovatelem forecastingu a poskytovatelem dat může být například zpřístupnění většího množství dat nebo lepšího vyhledávače mezi dokumenty.

4.2.3 Linked data

Linked data jsou součástí tzv. sémantického webu, což je ve zkratce rozšíření standardů World Wide Webu. Sémantický web je nový framework, který standardizuje formáty ukládání dat na webu a vyhledávání mezi nimi dle jasných pravidel. Cílem je umožnit počítačům provádět více úkonů a užitečné práce a vyvíjet systémy, které mohou podporovat důvěrné interakce napříč sítí [36].

Hlavním smyslem linked data je vytváření spojení či odkazů mezi daty z rozdílných zdrojů napříč celým World Wide Webem. Takto propojeny mohou být například dvě rozdílné databáze z různých organizací. Technicky se jedná o strukturovaná data zveřejněná ve strojově zpracovatelném formátu [37].

Článek „Quality Assessment for Linked Data: a Survey“ [38] se věnuje kvalitě linked data a identifikuje 18 metrik či parametrů, dle kterých lze tuto kvalitu posuzovat. V předchozích podsekcích bylo definováno již několik metrik, a ty byly vztaženy k tématu práce. Mnoho dalších metrik z článku je podobných již zmíněným metrikám nebo s tématem forecastingu vůbec nesouvisí, protože jsou silně navázány na linked data a nelze je upravit pro měření kvality zdrojů pro forecasting. Proto bude věnována pozornost pouze těm, které ještě nebyly prozkoumány.

- **Intenzita spojení mezi daty**

  Spojení mezi daty je jedním z hlavních charakteristik linked data, tvoří jejich podstatu, jak je z názvu patrné. Nicméně pro potřeby forecastingu jej lze interpretovat trochu v jiném smyslu. Lze předpokládat, že může určovat počet *citací* daného patentu nebo vědeckého článku, dokumentu. Citacemi jsou články spojené. Jedná se také o počet odkazů ukazujících na daný článek. Ověřování počtu citací je silný nástroj pro měření kvality článku, pro ověření pravdivosti článků. Čím více má článek citací, tím by měl být kvalitnější.

- **Škálovatelnost (Scalability)**

  Škálovatelnost je definovaná přesně jako zjištění, jestli čas na odpověď ze serveru na deset požadavků děleno deseti není delší než odpověď na jeden požadavek. Tato metrika je použitelná i pro zmiňovaný forecasting, kde bude muset server zpracovat velké množství požadavků, jelikož forecasting vychází z kvantitativních metod, tedy velkého množství informací.

- **Propustnost (Throughput)**

  Výkonnostní parametr propustnost indikuje kvalitu samotného technického řešení a infrastruktury zdroje. Propustnost zkoumá počet vyřešených požadavků za vteřinu, snahou služeb by mělo být tento počet maximalizovat. Propustnost jde ruku v ruce se škálovatelností. Požadavky, které posílá systém pro automatizovaný forecasting nových technologií, lze paralelizovat, a právě propustnost tak ovlivní rychlost celé predikce.

4.2.4 Další měření kvality dat

Téma kvality dat je obsáhlé samo o sobě i bez zaměření na konkrétní typ dat. Existuje tak mnoho publikací, které se zabývají obecnou kvalitou dat. Další články se zabývají kvalitou dat pro byznys a také zavádí nové metodiky pro jejich měření, jako například budiž uveden článek „AIMQ: a methodology for information quality assessment“ [39]. Jedná se sice o starší článek (z roku 2001), ale díky zavádění vlastních metodik pro vyčíslení metrik je pro tuto práci inspirativní. Také zmiňuje více než dvacet samostatných metrik, všechny relevantní z nich byly ale v této práci již diskutovány ve vztahu k automatizovanému forecastingu nových technologií.

4.3 Návrh měření kvality dat

V následujícím textu je na základě předchozích analýz, poznatků získaných v průběhu studia literatury a zkušeností autora, zadefinováno několik skupin a příslušných metrik spadajících do skupin tak, aby uspokojovaly potřeby pro měření kvalitu dat pro automatizovaný forecasting nových technologií. Toto

rozřazení je originální a objevuje se poprvé v této práci, stejně tak všechny metodiky (výpočty) pro kvantifikaci samotných metrik. Je to návrh, jak měřit kvalitu zdrojů pro automatizovaný forecasting nových technologií.

Při procházení identifikovaných metrik lze snadno nahlédnout, že jsou si některé více či méně podobné. Lze je logicky rozřadit do několika skupin. Při následném výpočtu kvality bude rozřazení výhodné, protože celé skupiny lze ohodnotit váhami podle důležitosti. Jedna skupina zastřešuje několik metrik a tuto strukturu je možné vizualizovat jako strom hloubky 1.

Dle logického rozčlenění budou definovány čtyři skupiny a popsány metriky do nich spadající. Těmito skupinami jsou: Trvanlivost, Technické zpracování, Hodnota informací a Nevyčíslené metriky. U každé metriky je definován výpočet, což je návod, jak kvantifikovat konkrétní metriku.

Všechny výpočty budou navrženy a normalizovány tak, aby nabývaly hodnot z intervalu 0 až 100 bodů. Kdykoliv hodnota výpočtu klesne pod 0, bude počítáno s nulou. Pokud hodnota přesáhne 100, bude počítáno s nejvyšší hodnotou, sto.

Použité pojmy

Systémem bude v následujícím textu myšlen systém pro automatizovaný forecasting nových technologií tak, jak je definován v podsekci 1.4.3.2.

Stahování stránky znamená přístup systému na frontendovou (část webu, která je dostupná běžnému uživateli) část, kdy stáhne celou HTML stránku a tu potom zpracovává, tedy extrahuje z ní potřebná data. Pro stahování byl použit program wget, je to známý program pro stahování stránek. Více informací lze nalézt v oficiální dokumentaci (dostupné na adrese www.gnu.org/software/wget).

Zdrojem jsou vždy myšleny datové zdroje pro automatizovaný forecasting technologií, jako jsou například ty popsané v předchozí kapitole. Poskytování datových zdrojů je nazýváno službou.

4.3.1 Trvanlivost (skupina)

Tato skupina pokrývá pouze jednu metriku.

4.3.1.1 Trvanlivost (metrika)

Náklady na napojení systému na zdroj nejsou zanedbatelné, jak je popsáno v kapitole 2. Proto je nutné ověřit, že zdroj bude fungovat ještě nějakou dobu po napojení. Jde v podstatě o předpoklad délky existence identifikovaného zdroje. Velkou roli hraje historie, po kterou již databáze existuje.

Výzkum z Mexika tvrdí, že průměrná délka životnosti podniku je 10 let [40]. Tento výzkum lze interpretovat několika způsoby, ale pro informační zdroje se autorovi jeví jako nejvhodnější předpokládat, že čím je zdroj starší,

tím je i stabilnější. Kvůli normalizaci na stobodovou stupnici ale budiž předpokládáno, že již překročení pěti let je významný milník a za každých pět let budou uděleny bonusové body. Na tom je založen také následující výpočet.

Výpočet

Proměnná *years* obsahuje počet let provozu zdroje, za každých pět let je hodnota násobena bonusem. Prvních pět let je bez bonusu za dlouhé trvání, ale druhých pět let je bonus dvojnásobek, dalších pět let trojnásobek a tak dále.

$$result = \sum_{i=1}^{years} \lceil \frac{i}{5} \rceil \tag{4.1}$$

Od výsledku budou odečteny negativní faktory, které mohou být příčinou odstavení zdroje z obchodních důvodů provozovatele (odečítaná hodnota je vždy psána s mínusem).

-5: Poskytovatel zdroje provozuje více podobných zdrojů

-10: Poskytovatel zdroje provozuje více podobných zdrojů a v obou lze nalézt stejné dokumenty

-20: Poskytovatel zdroje provozuje více podobných zdrojů a konkurují si, jsou zaměřeny na stejný nebo podobný okruh uživatelů

Samozřejmě se mohou vyskytnou skutečnosti, které nelze predikovat. Jako nákup jiného (druhého) zdroje společností, která provozuje (první) zdroj, na který je systém napojen. Důsledkem toho může být odstavení prvního zdroje. Takovou skutečnost je ale obtížné vyčíslit.

4.3.2 Technické zpracování

Skupina metrik s názvem „Technické zpracování“ se zabývá strojovou zpracovatelností poskytnutých dat. Systémy pro forecasting se potýkají při práci s daty s mnohými překážkami, které přidělávají analytikům práci a prodražují celý projekt. Pokud by se tyto překážky podařilo odstranit již na úrovni poskytovatele dat, systém by si převzal data jednoduše a nemusel by je upravovat, čistit a tvořit nad nimi operace za účelem správných predikcí.

Konkrétním příkladem je například poskytnutí dat přes API versus vybírání dat ze stažených stránek. Vhodně definované API zjednodušuje a zrychluje implementaci. Dalším příkladem budiž tvoření dotazů nad staženými daty – pokud by vyhledávač poskytl dostatečné možnosti při vyhledávání, jako je například řazení dle počtu citací, nemusel by systém provádět dotazy nad staženými daty, jednoduše by mu odpověď poskytl již sám vyhledávač.

4.3.2.1 Ochota spolupracovat

Někteří poskytovatelé dat jsou ochotni spolupracovat s jinými firmami či školami, a to buď zadarmo či za úplatu. Mohou být ochotni postoupit strukturovaná exportovaná data, otevřít API nebo dokonce zpřístupnit neveřejná data. Míru ochoty spolupracovat vyčíslí tato metrika.

Je nasnadě, že tato metrika je silně vztažená k samotnému provozovateli systému pro forecasting. Roli hraje jak silnou má vyjednávací pozici a jestli se jedná o jednotlivce či velkou firmu, přičemž velká firma bude mít pravděpodobně silnější schopnost poskytovatele dat přesvědčit k výhodnější spolupráci.

Výpočet

Iniciálně jsou poskytovatelé dat ohodnoceni 100 body. Body budou za každé narušení ochoty spolupracovat strženy či naopak přičteny za doplňkové služby. Způsoby ověření zájmu o vyjednávání o podmínkách jsou popsány v následující kapitole u vyčíslení této metriky.

100: Počáteční ohodnocení

-100: Poskytovatel nevyjednává o podmínkách spolupráce, žádné nejsou

-20: Nabídka na spolupráci není veřejná

-10: Služba je placená

+5: Za každou doplňkovou službu, kterou poskytovatel nabízí po domluvě či zaplacení nadstandardní služby a které zároveň přispěje ke zlepšení výkonu nebo účinnosti predikcí (například zvýšení výkonnosti serverů či zpřístupnění skrytých dokumentů)

4.3.2.2 Strojové zpracování

Existuje několik možností získávání dat ze zdroje. Jednou je stahování stránky, mnohem pohodlnější možností je stahování přes API. Hrozí tak menší riziko velkých změn části systému forecastingu pro stahování dat. Pokud se jen trochu změní stránka, ze které jsou data stahována, zpracování stažené stránky dat bude možná třeba celé přepracovat. Pokud se ale změní API, většinou se jedná o jednoduché, kosmetické úpravy a systém pro forecasting tomu lze za malé náklady přizpůsobit. Metrika měří kvalitu formátu dat ve vztahu ke zpracovatelnosti.

Otevřenému API ve standardu REST je udělena nejvyšší hodnota. V případě nedostupnosti API je nutné pro další zpracování stahovat celou stránku, což je hodnoceno méně body. Ještě více hodnocení zhorší dynamicky generované názvy tagů ve staženém dokumentu (HTML stránce), kvůli čemuž je mnohem náročnější strojově zpracovat staženou stránku.

Dynamické tagy jsou takové, jejichž jména se mění s každým obnovením stránky, tudíž není možné při strojovém zpracování přistupovat na ten stejný tag, který obsahuje potřebná data. Tuto taktiku někteří poskytovatelé dat záměrně užívají, aby ztížili strojové zpracování.

API může být k dispozici pouze po domluvě, tedy soukromě či placeně. V takovém případě bude zde ohodnoceno více body, ale v metrice „Ochota spolupracovat“ bude přiděleno bodů méně.

Výpočet

100: API ve standardu REST

90: API v libovolném standardu

70: Stahování dat pouze prostřednictvím exportování

50: Stažení HTML, bez dynamických tagů

-40: Dynamické tagy ve stažené stránce

-50: Ve stažené stránce nejsou očekávaná data (ta, která se zobrazují v prohlížeči), záměrné ztížení analýzy stažené stránky

4.3.2.3 Možnosti vyhledávání

Při vyhledávání není vždy cílem jen dostat co nejvíce výsledků. Někdy je potřeba konkrétněji určit, co chce systém vyhledávat. Tato metrika vyčíslí, jaké možnosti zpřesnění výsledků nabízí jednotlivé vyhledávače. Je mnoho možností, jak logiku vyhledávacího nástroje rozšířit – jak filtrovat výsledky a jak výsledky řadit.

Metrika odpovídá na otázku, co mohou vyhledávače dokumentů nabídnout a jak je ohodnotit. Za každou podporovanou funkcionalitu se přičítají body.

Rozšířené možnosti vyhledávače totiž zjednodušují implementaci samotného systému pro forecasting. Většinu z níže uvedených funkcionalit je totiž nutné pro forecasting implementovat a pokud by ji již zvládal sám vyhledávač, implementace systému by se tím zjednodušila.

Výpočet

Počáteční hodnota je stanovena na 0, za každou splněnou podmínku ze seznamu níže bude přičten daný počet bodů. Funkce jsou ohodnoceny body dle užitečnosti a míry využití v systému pro forecasting dle předchozí analýzy.

20: Vyhledání sousloví (například pomocí uvozovek)

20: Filtrování dle roku nebo rozsahu let

15: Řazení dle data

15: Definování slova, které se nemá vyskytovat ve výsledcích

15: Řazení dle relevance ve vztahu k vyhledávanému výrazu

15: Řazení dle počtu citací

10: Změna počtu výsledků na stránku (na stránce může být více než 100 výsledků)

10: Výběr typu dokumentu (v případě, že vyhledává mezi více typy)

5: Výběr jazyka výsledků (například dokumenty psány pouze v angličtině)

5: Jméno autora / autorů

5: Země původu dokumentu

Je možné, že vyhledávač bude obsahovat další unikátní a užitečné funkce. V takovém případě by měly být ohodnoceny právě 0 až 30 body dle míry užitečnosti v systému pro forecasting. Je třeba vzít v potaz, jak často bude funkce využita a také jak výrazně usnadní výpočet predikce.

4.3.3 Hodnota informací

Tato skupina metrik se zabývá hodnotou informací ve zdrojích. Jejím cílem je zkoumat dokumenty ve zdrojích obsažené a na základě toho určit, jestli jsou vůbec vhodné pro samotný forecasting. Zdali zdroje obsahují dostatek dokumentů a zdali jsou ty dokumenty relevantní a vyhledatelné.

4.3.3.1 Objem dat

Počty dokumentů obsažených v databázích se pohybují v miliónech. Čím více bude mít databáze záznamů, tím větší bude šance nalézt cenné informace. Na druhou stranu se zvýší čas zpracování. Důležitou roli ale hrají vznikající technologie, tudíž všechny dokumenty mladší deseti let.

Růst zájmu o technologií lze dle analýzy detekovat zhruba po zkoumání časového horizontu tří až pěti let. Deset let je bezpečný odhad v případě, že je o technologii pozvolnější nárůst zájmu, aby byl v kontextu času stále detekovatelný.

Čím více bude zdroj obsahovat dokumentů za posledních deset let, tím pravděpodobněji se mezi nimi budou nacházet možné zajímavé informace.

Aby byl počet normalizovaný na stobodovou stupnici, bude počet nových dokumentů za posledních deset let vydělen deseti tisíci. K tomuto výpočtu bylo přikročeno na základě výpočtů z následující kapitoly a článku „Global scientific output doubles every nine years“ [41], který zkoumá celkový objem vědecké práce a její nárůst.

Výpočet

Výsledek bude číslo zaokrouhlené na jednotky. *documents_10_years* je počet dokumentů za posledních deset celých let. Pro rok 2018 jsou to dokumenty za rok 2008 až 2017 včetně.

Do vyhledávače často nelze zadat prázdný dotaz tak, aby byly vyhledány všechny články, proto je doporučeno zadat slovo „technology“, případně technologie pro české vyhledávače. Tento test lze provést jakýmkoliv výrazem, je doporučeno užít výraz, na který se bude forecasting zaměřovat. Pro obecné účely dobře poslouží „technology“.

$$result = \frac{documents\_10\_years}{10000} \tag{4.2}$$

4.3.3.2 Reputace, úroveň dat

Ať už databáze obsahuje desítky miliónů záznamů, takový objem nebude k užitku za předpokladu, že záznamy jsou nevalné technické kvality. Takovou mohou vykazovat například nevědecké nerecenzované články. Pro výpočet celkové kvality je nezbytné ohodnotit reputaci a kvalitu zdroje.

Je třeba zavést přesně specifikované kategorie, do kterých budou jednotlivé zdroje spadat. Slovo „agregované“ udává, že dokumenty nepochází přímo od datového zdroje (například je nezveřejňuje sama univerzita či studijní skupina, nebyly vloženy k recenzi tomuto zdroji), ale jsou sbírány a indexovány.

Výpočet

100: Recenzované články, patenty

80: Agregované a recenzované články, patenty

60: Agregované a recenzované články, patenty a knihy

40: Agregované články, patenty, knihy

20: Agregované články, patenty, knihy, zprávy

0: Další zdroje, vkládání dokumentů nepodléhá recenzování ani kontrole

4.3.3.3 Potenciál zdroje pro predikci

Hlavním předpokladem pro fungování predikcí nových technologií je, že zdroj obsahuje dokumenty, které se o nové technologii zmiňují ještě předtím, než je známá veřejnosti. Pojem „technologie známá veřejnosti“ nemá přesnou definici, ale pro vyčíslení této metriky je nutné takovou definici zavést.

Hype cycle je produkt společnosti Gartner přesně popsaný v podsekci 2.2.3. Pro připomenutí – je to křivka, která zobrazuje mimo jiné i technologie na počátku zájmu, ve fázi počátku zájmu o technologii. Je zveřejňována jednou

ročně a zveřejněním se informace o novinkách dostávají mezi širokou veřejnost. Tím se technologie stává známá veřejnosti. Právě technologie před zveřejněním v Hype cycle lze považovat za vznikající.

Tato metrika zkoumá, jestli dokumenty obsažené v databázích opravdu reflektují následný historický vývoj. Tedy jestli databáze opravdu obsahují dokumenty se zmínkou o vznikajících technologiích, které se v následujících letech objeví například na počátku Hype cycle jako novinky.

Pravdivost zdrojů je pro měření kvality dat často užívaná metrika, je definována u big data, linked data i pro obecné měření kvality dat. Často je ověřována citacemi, to ale není pro predikci technologií tak vhodné, zabývá se totiž kvantitativní analýzou a citace odráží spíše kvalitu. Pravdivost dokumentů bude reflektovat právě metrika Potenciál zdroje pro predikce. Ta je založena na množství dokumentů věnující se jedné technologii, proto bude měřeno množství dokumentů.

Pro forecasting nových technologií lze pravdivost definovat právě tak, jestli zdroj obsahoval dokumenty, které byly schopny předpovědět nárůst zájmu o technologii. Jinými slovy je to možnost, jestli zdroj mohl technologii korektně předpovědět.

Zdroje budou bodovány na základě následujících navržených testů.

Výpočet

Aby test pokrýval obecné téma technologií, jsou vybrány dvě technologie a konkrétní roky, kdy se objevily na Hype cycle. Pokud se systém pro forecasting nových technologií má věnovat konkrétní oblasti, je doporučeno vyhledat jiné výrazy (term), které by seděly záměru právě zkoumané oblasti. Níže provedené výpočty na konkrétních technologiích jsou tedy příkladem, na kterém je vysvětleno fungování této metriky. Jsou použitelné v případě, že bude zkoumána oblast obecně nových technologií.

Pro ilustraci výpočtu jsou vybrány dvě technologie, budou vyhledány dva výrazy (term) – autonomní vozidla a 4D tisk.

- **autonomous vehicle – 2010**

  Technologie autonomních vozidel se poprvé na Hype cycle objevila v roce 2010. Všechny zmínky, které jsou obsaženy v databázích a jsou starší roku 2010, budou sečteny. Přímo graf z roku 2010 je k dispozici k nahlédnutí v podsekci 2.2.3.

- **4D printing – 2016**

  Technologie 4D tisku se v Hype cycle poprvé objevila roku 2016. Nová dimenze značí transformaci v čase, ale pro účely měření potenciálu zdroje pro predikce je lhostejné, čeho se tato technologie přesně týká. Klíčové je identifikovat dokumenty mladší roku 2016 obsahující zmínku o ní.

Samotný výpočet by měl být relativní vzhledem k počtu celkových záznamů o dané technologii ve zdroji. V obecném případě lze zkoumat více než dvě zvolené technologie, vzorec bude zapsán v obecné formě. *tech_all_years* je celkový počet dokumentů se zmíněnou technologií a *tech_before_hype* je počet dokumentů před rokem, ve kterém se technologie objevila s Hype cycle. *term* je počet výrazů, které do vyhledávání vstupují, zde by to byly dvě – autonomní vozidla a 4D tisk.

$$result = \sum_{i=1}^{term} \lceil (\frac{tech\_before\_hype_i * 100}{tech\_all\_years_i}) * (\frac{1}{term}) \rceil \tag{4.3}$$

$$result = \sum_{i=1}^{term} \lceil \frac{tech\_before\_hype_i * 100}{tech\_all\_years_i * term} \rceil \tag{4.4}$$

Celkovou hodnotou (*result*) je průměrné procento záznamů o nových technologiích.

Název proměnné *tech_before_hype* bude v následující kapitole zkrácen na *tech_b_hype*.

4.3.3.4 Relevance vyhledaných položek k dotazu

Tato metrika kontroluje, jak přesný je vyhledávač pro identifikované zdroje, jak velkou shodu nabízí, jak precizně vyhledává. Pro evaluaci je navržen podobný test jako pro Potenciál zdroje pro predikce. Tyto dvě metriky spolu totiž souvisí. Vyhledávač může sice vrátit mnoho výsledků, ale některé nemusí být vůbec relevantní. Konkrétní hodnoty budou vypočteny a vysvětleny v následující kapitole.

Výpočet

Výpočet má být proveden na stejných výrazech, které byly zvoleny v předchozí metrice, zde jsou to pro demonstraci autonomní vozidla a 4D tisk. Výpočet se provede manuální kontrolou prvních deseti navrácených dokumentů. Za každý dokument správně spojený s technologií je přidáno deset bodů.

V obecném případě je možné kontrolovat více dokumentů a vzorec tomu přizpůsobit. Ale již při kontrole prvních deset lze získat dobrou představu, jestli vyhledávač vrací relevantní výsledky či ne.

Některé technologie mohou mít názvy podobné jiným, ale vyhledávač by měl správně detekovat ty, které pravděpodobně zažívají nárůst a nestavět na první místa ty, které s nárůstem zájmu nesouvisí. Je tedy možné, že mezi autonomními vozy se objeví studie o vozidle na dálkové ovládání, na druhou stranu konkrétní příklad je k dispozici v následující kapitole o vyčíslení metrik.

- **autonomous vehicle – 2010**

  Mnoho dokumentů se nemusí týkat konkrétně technologie samořiditelného vozu. Jsou mezi nimi patenty pro navigaci, ohyb zrcátek a podobné.

Ty jsou v pořádku, budou ohodnoceny deseti body. Nevhodné jsou ale patenty týkající se aut na dálkové ovládání. Tyto jsou zavádějící a neměly by se objevovat ve vyhledávačích na prvních místech.

- **4D printing – 2016**

  Při vyhledání 4D tisku by neměly být dokumenty nalezené na prvním místě věnovány pouze 3D tisku.

Vzorec je vyjádřen pro obecný počet výrazů (*term*) a obecný počet kontrolovaných dokumentů (*docs*). *doc_checked* je kontrolovaný dokument, který dle počtu kontrolovaných dokumentů lze ohodnotit 0 nebo $\frac{100}{docs}$ body.

$$result = \sum_{i=1}^{term}(\sum_{j=1}^{docs} doc\_checked_j) * (\frac{1}{term}) \tag{4.5}$$

$$result = \sum_{i=1}^{term}\sum_{j=1}^{docs} \frac{doc\_checked_j}{term} \tag{4.6}$$

4.3.4 Nevyčíslené metriky

Poslední částí jsou Nevyčíslené metriky, což ovšem není plnohodnotná skupina, jelikož neobsahuje metriky, pro které by byl zaveden výpočet, vyčíslení. Jedná se o metriky, které je užitečné zmínit a případně podrobit dalšímu zkoumání, zdali mají vliv na systém pro forecasting. Výpočty nebyly zavedeny z různých důvodů popsaných u samotných metrik. Na druhou stranu se může jednat o metriky, u nichž naplnění jejich požadavků je nutná podmínka pro užití zdroje. Je to tedy binární kritérium, pokud zdroj podmínku nesplňuje, tak ho není vhodné užít pro automatický forecasting nových technologií.

4.3.4.1 Korelace s expertními odhady

Velmi užitečné by bylo porovnávat informace ze zdrojů s názory expertů na danou doménu. Uskutečnitelné by to bylo například dotazníkovým šetřením. Bylo by možné dokumenty vzešlé z testů v části Relevance vyhledaných položek k dotazu prezentovat vybraným expertům, kteří by jejich relevanci ohodnotili. Samozřejmě se stále jedná o subjektivní metriku, čemuž je snaha se v exaktní a objektivní kvantifikaci vyhnout, ale bylo by zajímavé pozorovat konfrontaci expertů s výsledky vyhledávání.

4.3.4.2 Dostupnost a výkon

Pokud je nějaký zdroj zvažován pro použití v automatizovaném forecastingu, je potřeba, aby byl k dispozici, kdykoliv je nutné generovat predikce. Zdroj by neměl mít časté výpadky. V nynější době je u profesionálních služeb běžně

k vidění dostupnost přesahující 99,5 %, což je pro většinu zákazníků dostačující. Služba je dostupná, když je operabilní, plní svůj účel a funguje. Výpadky jsou problém, když poběží několikahodinový výpočet predikce, systém pro forecasting nesmí spadnout a přerušit hodiny výpočtu, zejména musí dostat potřebná data.

Pro známé zdroje provozované velkými firmami by s dostupností neměl být problém. Tato metrika je zmíněna pro úplnost.

Výkon souvisí se škálovatelností a propustností tak, jak jsou popsány u linked data v podsekci 4.2.3. V dnešní době, stejně jako s dostupností, ubývá problémů s řízením výkonu služeb. Díky škálovatelným řešením infrastruktury jsou často nabízeny plány, kdy si zákazník připlatí, a za to jednoduše dostane vyšší výkon. Pro získávání dat ve forecastingu nových technologií, které trvají hodiny, je výkon v dnešní době spíše na druhém místě zájmu. Měření výkonu samo o sobě je již dobře definované, je to standardizovaná disciplína, tato práce se spíše věnuje specifikům forecastingu než specifikům požadavků na infrastrukturu, ty není třeba znovu zadefinovávat.

4.3.4.3 Aktualita

Při analýze budoucích trendů jsou potřeba aktuální dokumenty, a to z posledních přibližně deseti let. Blíže se tomuto odhadu věnuje metrika Objem dat v části 4.3.3.1. Požadavek aktuality odpovídá na otázku, zdali obsahuje databáze nejnovější dokumenty.

Jedná se o jednoduché ověření, které má smysl provést vždy, než začne práce se zdrojem. Je třeba ověřit, jestli je stále aktualizovaný. Nemá smysl přidávat tuto metriku do vyčíslitelných parametrů, protože pro forecasting nových technologií je nevhodné pracovat s neaktuálními daty, potažmo zdroji.

4.3.4.4 Legislativní omezení

Metrika zkoumá, jestli má jednotlivec či potenciální provozovatel systému pro forecasting právo užít data pro svůj záměr. Metrika je relevantní zejména při plánování projektu a její nesplnění může vést k nepříjemným legislativním důsledkům.

4.3.4.5 Metadata

Metadata jsou jednoduše řečeno data o datech. V kontextu diskutovaných dokumentů lze metadata chápat konkrétně jako autor dokumentu, rok vzniku, abstrakt (či náhled knihy, perex zprávy), počet citací.

Výše zmíněná metadata jsou pro forecasting nových technologií důležitá. V případě, ze zdroj tato data neposkytuje, není vhodným kandidátem jako zdroj pro predikci nových technologií. Stejně jako Aktualita a Legislativní omezení jsou i Metadata metrikou, jejíž splněním je třeba výběr zdroje podmínit.

Kapitola 5

Výpočet hodnot metrik

V kapitole 3 byly identifikovány reálné zdroje, v kapitole 4 byla navržena metodologie měření jejich kvality pomocí metrik a v této kapitole jsou navržené metriky vyčísleny, tedy jsou provedeny navržené výpočty.

Tato kapitola slouží také k ověření navržených metodik výpočtů v tom smyslu, že je možné výpočty provést a ohodnotit tak zdroje. V opačném případě by navržené metriky nebyly použitelné. Na základě výpočtů v této kapitole také byly metriky upraveny do aktuální podoby, tento postup je popsán v sekci 5.6 na konci kapitoly.

V případě, že některé skutečnosti nelze ověřit přesně, bude proveden odhad hodnoty. U každého odhadu bude popsán myšlenkový postup, aby bylo jasné, jak lze odhady provádět.

Všechny zjištěné skutečnosti vyplývají z oficiálních materiálů jednotlivých zdrojů, pokud není uvedeno jinak.

Kompletní přehled naměřených hodnot v jedné tabulce pro všechny vyčíslitelné metriky pro zkoumané zdroje je k dipozici na konci této kapitoly.

5.1 Výběr zdrojů k ohodnocení

Za každou ze tří skupin zdrojů – patentové zdroje, zdroje s odbornými znalostmi, zprávy – je vybráno několik reprezentantů a na nich jsou demonstrovány výpočty metrik. Zdroje, na kterých jsou výpočty předvedeny, jsou vybrány tak, aby byly výsledné kvality co možná nejrůznější. To znamená, že ze dvou zdrojů, které by si mohly být na první pohled podobné, bude vybrán právě jeden. Zároveň nejsou vybrány zdroje, ke kterým je nutné vytvořit účet a zaplatit poplatek za přístup.

Patentové zdroje

Patentové dokumenty jsou jednou z nejcennějších zdrojů pro automatizovaný forecasting nových technologií. Vybráni byli tři největší zástupci, ke kterým měl autor přístup. K Derwent World Patents Index se nebylo možno

dostat přes autentizaci i přes to, že autor odeslal žádost a vstup z akademických důvodů. Po týdnu čekání na odpověď bylo rozhodnuto tento zdroj z hodnocení vynechat.

Zdroje s odbornými znalostmi

Pro zdroje s odbornými znalostmi byly záměrně vybrány více než tři zdroje. Google Scholar je obsahově největší ale také nejuzavřenější, bude zajímavé vidět jeho porovnání s Scopus a Web of Science, což jsou velcí hráči na poli poskytování odborných bibliografických dat. Dalším k prozkoumání je Summon ČVUT, který má ještě jiný obchodní model než předchozí zmíněné databáze, bude zajímavé vidět jejich porovnání.

Zprávy

Pro zprávy byli vybráni Google News a Bing News, což jsou dva velcí konkurenti, a jeden zástupce z Česka – Technický týdeník, který by mohl nabídnout zajímavou alternativu ke světovému standardu. Reuters nebylo vybráno, jelikož je velmi podobné Google News a Bing News a nepřineslo by tak do závěrů žádné nové prvky.

5.2 Trvanlivost

Pro výpočet trvanlivosti stačí dosadit do vzorce 4.1 rozdíl aktuálního roku (2018) a roku vzniku (z tabulky 3.1 na konci kapitoly 3). Goole Scholar dostane o deset bodů méně, protože mezi jeho dokumenty se objevují i dokumenty z Google Books a oba zdroje jsou provozovány jedním provozovatelem, za tento fakt se dle pravidel odečtou body. Stejně tak Scopus dostane o pět bodů méně, jelikož Elsevier provozuje podobnou databázi – Engineering Village.

Ohodnocení zdrojů metrikou Trvanlivosti

Název zdroje	Trvanlivost
Espacenet	50
Patentscope	24
Google Patents	21
Google Scholar	17
Summon ČVUT	5
Scopus	60
Web of Science	24
Google News	34
Bing News	13
Technický týdeník	18

Tabulka 5.1: Ohodnocení zdrojů metrikou Trvanlivosti

5.3 Technické zpracování

Patentové zdroje

Espacenet, potažmo Open Patent Services, poskytuje službu zdarma omezenou velikostí stahovaných dat za týden a po platbě nabízí množství doplňkových služeb, cena je veřejná. Nabízí API ve standardu REST. Espacenet si dává záležet na implementaci funkcí do svého vyhledávání, přesto ale chybí více možností řazení kromě základního řazení dle času. Celkem byly napočítány 4 užitečné funkce (omezení času (20 bodů), vyhledání slovního spojení (20 b.), klasifikace dle kategorie patentu (10 b.), výběr jazyka patentu (5 b.)).

Patentscope má veřejnou nabídku na spolupráci s udanou cenou, jako bonus nabízí abstrakty ve francouzštině a vyfocenou patentovou přihlášku, tyto výhody jsou ale pro zkoumané užití zbytečné. API obsahuje všechny potřebné parametry, ale není ve standardu REST. Vyhledávání podporuje určení vlastního časového rozsahu (20 bodů), definici jazyka patentu (5 b.), rozřazení patentů do kategorií a podkategorií (10 b.), více možností řazení (10 b.) a upravení počtu patentů na seznam (10 b.).

Google má ve většině svých služeb (Scholar, Patents, News, klasické vyhledávání) API ve stavu „deprecated“, což znamená, že existovalo, ale nyní je nedostupné. Výjimkou jsou Google Books.

Ostatní zdroje s poskytnutím API většinou problém nemají. Nebyly nalezeny zmínky o zájmu Google Patents o spolupráci. Strojové zpracování je pokryto stažení informací o vyhledaných dokumentech ve formátu CSV (Comma-Separated Values), tedy export dat. Google Patents má nejpropracovanější vyhledávání, implementuje všechny relevantní podmínky popsané u této metriky.

Zdroje s odbornými znalostmi

Google Scholar nemá veřejné API ani nemá zájem o spolupráci. Má inteligentní vyhledávání, které obsahuje mnoho podobných funkcí jako Google Patents (vyhledání spojených výrazů (20 b.), zakázání výskytu slov (15 b.), řazení dle data (15 b.), relevance (15 b.), označení rozmezí pro roky (10 b.)). Stránku lze bez problémů stáhnout a extrahovat z ní potřebná data, tagy jsou statické.

Vyhledávač Summon ČVUT má velmi pokročilé možnosti vyhledávání, nikde ale nebyl nalezen odkaz na API ani na možnosti spolupráce. Z toho důvodu byl provozovatel kontaktován autorem práce s požadavkem o vysvětlení provozu API a nastínění ochoty ke spolupráci. Do doby odpovědi bude API i ochota spolupracovat ohodnoceny 0. Summon je ale standardizované řešení a na stránkách poskytovatele popisuje API ve standardu REST, mělo by tedy být k dispozici. Summon ztěžuje analýzu po stažení dynamickými tagy a navíc stránka neobsahuje očekávaná data, tedy ta, který se zobrazí v prohlížeči. Vyhledávání má Summon velice propracované – obsahuje více možností řazení (dle relevance za 15 b. A dle data za 15 b.), definice přesného výrazu (20 b.),

definice rozmezí dat (20 b.), kategorií dokumentů (10 b.), vlastní definice vyhledávaných výrazů za pomoci logických spojek AND, OR, NOT (10 b.) a možnost zobrazit pouze recenzované či akademické dokumenty (10 b.).

Scopus má velice dobře řešené API i s dokumentací, na druhou stranu je to placená služba se skrytou nabídkou. K vyhledání je možné definovat typ dokumentu (10 b.), rozsah let (20 b.), přesné výrazy (20 b.), klíčová slova (10 b.) a přesné místo vyhledávání (v abstraktech, v klíčových slovech) (10 b.).

Web of Science nabízí profesionální API se všemi potřebnými parametry, přičemž nabídka není veřejná. Do služby se lze přihlásit ČVUT účtem a procházet tak záznamy. Web of Science obsahuje pokročilé vyhledávání s možnostmi využití logických operátorů stejně jako Summon (10 b.), přesné výrazy (20 b.), rozsah let (20 b.), výběr z kategorií dokumentů (10 b.), typů dokumentů (10 b.), více možností řazení (dle relevance za 15 b. A dle data za 15 b.) a výběr jazyka dokumentu (5 b.).

Ohodnocení zdrojů metrikami Technického zpracování

Název zdroje	**Ochota spolupracovat**	**Strojové zpracování**	**Možnosti vyhledávání**
Espacenet	100	100	55
Patentscope	90	90	55
Google Patents	0	70	100
Google Scholar	0	50	75
Summon ČVUT	0	0	100
Scopus	80	100	70
Web of Science	80	100	100
Google News	10	50	40
Bing News	90	100	40
Technický týdeník	0	50	10

Tabulka 5.2: Ohodnocení zdrojů metrikami Technického zpracování

Zprávy

Google News API je ve stavu „deprecated“, nelze jej užít. Ve stažené stránce se nevyskytují dynamické tagy. Vyhledávat lze dle přesného výrazu (20 b.) i bez konkrétních slov (15 b.), v definovaném jazyce (5 b.).

Bing News má veřejnou nabídku na zpřístupnění API, cenové stupně se odvíjí podle počtu dotazů ze serveru. Vrací ale zprávy nejdále za posledních třicet dní. Možnosti vyhledávání jsou stejné jako u Google News.

Technický týdeník nemá k dispozici API, nezabraňuje ale žádnými metodami ve stahování a čtení HTML stránky. Do vyhledávání lze zadat pouze konkrétní rok vydání (20 body je ohodnoceno rozmezí, omezení na jeden rok bude ohodnoceno polovinou bodů), žádné další užitečné funkce nenabízí.

Žádný ze zdrojů nenabízí pohodlné filtrování dle rozsahu let. Technický týdeník nabízí možnosti filtrování po jednom roce, čímž alespoň poskytuje nějakou možnost filtrování, Google News a Bing News ale dovedou vrátit pouze dokumenty z poslední doby (den, týden, měsíc).

5.4 Hodnota informací

V tabulkách 5.3 jsou zapsány počty nalezených a vyhodnocených patentových dokumentů, 5.4 zdrojů s odbornými znalostmi a v tabulce 5.5 zpráv ve vztahu k metrikám Potenciálu zdroje pro predikce a Relevanci dotazů. Jsou vždy pro autonomní vozidla (av) a pro 4D tisk (4D).

Patentové zdroje

Espacenet neposkytuje přesné číslo o počtu článků, pouze pokud je jich na 10 000, zobrazí informaci, že článků je nad 10 000. To je ale nepoužitelné, proto je nutné udělat kvalifikovaný odhad. Byly zkoumány intervaly po jednom měsíci v několika rocích a průměrně je publikováno vždy 5 000 dokumentů. To znamená průměrně 60 000 za rok. Espacenet agreguje pouze patenty, dostává nejvyšší hodnocení za reputaci. Co se týče Potenciálu zdroje pro predikce, obsahuje 7 308 záznamů o autonomních vozidlech, z toho 1 364 před rokem 2010. O 4D tisku obsahuje 97 zmínek, z toho 65 je před rokem 2016. Relevantních zmínek o autonomních vozidlech je z prvních deseti polovina a o 4D tisku nula. Ono „4D“ většinou sloužilo k popisu obrázku 4D a náhodně bylo spojeno s patentem týkajícím se tisku.

Patentscope zobrazuje přesný počet dokumentů. Pro zjištění jejich počtu za posledních deset let byl použit následující dotaz: „FP:technology AND PD:([01.01.2008 TO 31.12.2017])“. Použitý dotaz je uveden tam, kde je nutné jej psát manuálně.

Pro zjištění objemu dat v Google Patents byl použit dotaz „(technology) before:priority:20171231 after:priority:20080101“. Počet je přes tři milióny, což překračuje bodovou hranici a tudíž je ohodnoceno sto body.

Zdroj a výraz	**tech_all_years**	**tech_b_hype**	**doc_checked**
Espacenet – av	7308	1364	50
Espacenet – 4D	87	65	0
Patentscope – av	6306	1180	40
Patentscope – 4D	1011	984	0
Google Patents – av	411792	238349	100
Google Patents – 4D	49	46	0

Tabulka 5.3: Hodnoty proměnných z patentových zdrojů vstupujících do výpočtů metrik Potenciálu zdroje pro predikce a Relevance k dotazu

Důležité je držet se zvoleného řazení pro všechny zdroje – preferované je dle relevance. Tedy nalezne dokumenty, které by měly být co nejblíže vyhledávanému výrazu.

Zdroje s odbornými znalostmi

Google Scholar obsahuje 640 000 dokumentů zaměřených na technologie za posledních deset let.

Summon obsahuje přes 1 000 000 dokumentů zaměřených na technologie, dostává plný počet bodů za obsah.

Scopus indexuje přibližně 1 500 000 dokumentů za posledních deset let s tématem technologie. Pro vstup do databáze byl užit účet ČVUT. Reputace se může lišit od 20 až do 60, jelikož lze dohledat novinové články, které spadají do nižší kategorie hodnocení. Na druhou stranu je lze vyfiltrovat a Summon je řazen do kategorie Odborných zdrojů, je tak posuzován a při vyhledávání jsou novinové články vždy filtrovány.

Web of Science obsahuje 952 000 článků za posledních deset let na téma technologie. Pro vstup do databáze byl užit účet ČVUT.

Zdroj a výraz	**tech_all_years**	**tech_b_hype**	**doc_checked**
Google Scholar – av	2120000	1060000	70
Google Scholar – 4D	610000	299000	20
Summon ČVUT – av	105835	47116	100
Summon ČVUT – 4D	15127	13720	20
Scopus – av	32990	13252	70
Scopus – 4D	34	17	30
Web of Science – av	19916	7665	50
Web of Science – 4D	169	49	70

Tabulka 5.4: Hodnoty proměnných ze zdrojů s odbornými znalostmi vstupujících do výpočtů metrik Potenciálu zdroje pro predikce a Relevance k dotazu

Zprávy

Počet zpráv za posledních deset let se pro Google News se nepodařilo zjistit, jelikož nepodporuje vyhledávání pro konkrétní roky. Při celkové velikosti 350 miliónů dokumentů pravděpodobně bude dostatečně pokrývat požadavek na obsah.

Bing News také nedostatečně prezentuje množství nalezených článků. Odhad není možné provést.

Technický týdeník obsahuje 3650 článků obsahující slovo technologie. Je to skoro čtvrtina jeho celkového objemu dat. Co se týče kvality dat obsahuje nejkvalitnější data ze skupiny Zprávy, jelikož příspěvky jsou kontrolované. Oproti tomu Google News a Bing News mohou obsahovat i blogové příspěvky, které nepodléhají žádné kontrole.

Právě kvůli obchodnímu modelu, kdy Google News a Bing News vrací pouze zprávy z poslední doby (den, měsíc), nelze provést navržené testy.

Zdroj a výraz	**tech_all_years**	**tech_b_hype**	**doc_checked**
Google News	–	–	–
Bing News	–	–	–
Technický Týdeník – av	550	80	30
Technický Týdeník – 4D	240	150	0

Tabulka 5.5: Hodnoty proměnných ze zpráv vstupujících do výpočtů metrik Potenciálu zdroje pro predikce a Relevance k dotazu

Ohodnocení zdrojů metrikami Hodnoty informací

Název zdroje	**Objem dat**	**Reputace, úroveň dat**	**Potenciál zdroje pro predikci**	**Relevance vyhledaných položek k dotazu**
Espacenet	60	100	44	25
Patentscope	24	100	59	20
Google Patents	100	100	76	50
Google Scholar	64	60	50	45
Summon ČVUT	100	60	69	60
Scopus	100	80	46	50
Web of Science	95	80	35	60
Google News	100	0	–	–
Bing News	–	0	–	–
Technický týdeník	1	20	40	15

Tabulka 5.6: Ohodnocení zdrojů metrikami Hodnoty informací

5.5 Nevyčíslené metriky

Je zapotřebí ověřit, že zdroje splňují povinné metriky – Aktualitu, Legislativní omezení a Metadata. Na druhou stranu metriky Korelace s expertními odhady a Dostupnost a výkon jsou dostatečně diskutovány již u návrhu v kapitole 4 a není nutné jim již v této části věnovat další prostor.

Aktualita

Pro všechny zdroje byl proveden test na Objem dat a Potenciál zdroje pro predikce. Při těchto testech bylo také kontrolováno, že zdroj obsahuje data z aktuálního roku.

Legislativní omezení

Výrazné legislativní omezení nebyly nalezeny. Jedno z nejpřísnějších omezení má Google Scholar, který ve svých podmínkách užití deklaruje, že si nepřeje, aby jeho stránky byly využity jinak, než je přesně popsáno v návodu (podmínky jsou dostupné na adrese www.google.com/intl/en/policies/terms). Trestem za porušení předpisů je znepřístupnění služby. Dle zkušeností autora Google detekuje opravdu přítomnost robotů, kteří stahují stránky, a blokuje jim přístup.

Metadata

Všechny zdroje obsahují důležité parametry, jako jsou abstrakt, shrnutí nebo perex (v případě zpráv) a rok vzniku. Problematická je absence filtrování zpráv dle roku u Google News, což způsobuje, že je tato služba pro diskutovaný forecasting nepoužitelná.

5.6 Dopady vyčíslení metrik

Na základě vyčíslení metrik byly posouzeny jejich dopady pro výpočet kvality zdrojů, z čehož byly vyvozeny níže popsané závěry.

5.6.1 Korekce metrik

V této sekci je popsáno, jak byly na základě výpočtů korigovány a měněny metriky. Návrh nebyl pouze jednorázový v tom smyslu, že by metriky byly navrženy a vyčísleny. Návrh byl několikrát revidován právě na základě vyčíslení z této kapitoly. Níže jsou popsány nejvýznamnější provedené změny. Ostatní změny se většinou týkaly změn vzorců pro výpočet a drobnějších změn v definicích.

API

Z Technického zpracování byla odstraněna metrika „API“. Metrika původně udávala kvalitu API ve smyslu obsahu, tedy jestli API obsahovalo všechny parametry. Ukázalo se, že při vysoké hodnotě metriky Ochota spolupracovat bylo vysoké i hodnocení API, stejně tak při nízké bylo nízké i API. Tato vysoká korelace se vyskytla u všech hodnocených zdrojů, proto byla metrika API odstraněna jako redundantní a byla začleněna do metriky Ochota spolupracovat.

Různorodost druhotných zdrojů

V celé práci je slovem „zdroj“ označován právě poskytovatel dat pro forecasting, jako například databáze Scopus. Tento zdroj má i své zdroje, ze kterých čerpá a jejichž dokumenty agreguje. Takové zdroje budou nazývány

druhotné zdroje. Tedy zdroje mají druhotné zdroje, ze kterých čerpají dokumenty. Například zdroj Google Scholar má mezi svými druhotnými zdroji ScienceDirect, Springer a desetitisíce dalších stránek, časopisů, databází a vydavatelství.

Tato metrika vyjadřovala, z kolika různých druhotných zdrojů pochází data. Pro forecasting má malá různorodost negativní dopady, protože je potřeba zkoumat zvyšující se zájem napříč výzkumnými týmy, a ne pouze na jedné škole či v jednom časopise. Díky rozhledu do celosvětové činnosti výzkumníků se lépe měří nárůst zájmu.

Ukázalo se ale, že druhotných zdrojů jsou desítky tisíc, ať už jsou to organizace nebo univerzity pro patenty nebo zpravodajské agentury a webové stránky pro zprávy. V této metrice byly všechny hodnocené zdroje velice silné a ničím je neodlišovaly, metrika proto byla zrušena.

Výkon

Původně byly definovány metriky pro měření výkonu. Po zvážení autora byly ale přesunuty do Nevyčíslených metrik, jelikož zadefinováním těchto metrik byly duplikovány obecně platné metriky pro měření výkonnosti a nepřinášely pro práci žádnou přidanou hodnotu. Dalším důvodem byla těžká ověřitelnost a výpočty, jelikož většina zdrojů vyžaduje platby za přístup k API a až po tomto procesu by bylo možné vůbec začít výkon testovat. Jelikož autor kontaktoval několik zdrojů se žádostmi buď o autorizaci k přístupu do zdroje nebo je žádal o odpověď na dotaz a nedostal žádné odpovědi, bylo by pravděpodobně testování všech zdrojů jednoduše neproveditelné.

5.6.2 Zprávy

Pro zdroje typu zpráva nevyšlo ohodnocení pozitivně. Největší problém je nemožnost vyhledávání zpráv dle roku. Pro studovaný forecasting je ale zkoumání zpráv z konkrétních let klíčové, jelikož bez toho není možno zkoumat nárůst zájmu ani sledovat trendy dále, než je povolený limit. Limity bývají den, týden nebo měsíc. Pro automatizovaný forecasting nových technologií se tak zdroje typu zpráv jeví jako nevhodné.

5.7 Přehled hodnot všech metrik

Následující tabulka 5.7 přehledně shrnuje všechny hodnoty metrik vypočtené pro jednotlivé zdroje v rámci této kapitoly.

Tabulka 5.7: Souhrnná tabulka hodnot metrik

Název zdroje	Trvanlivost	Ochota spolupracovat	Strojové zpracování	Možnosti vyhledávání	Objem dat	Reputace, úroveň dat	Potenciál zdroje pro predikci	Relevance vyhledaných položek k dotazu
Espacenet	50	100	100	55	60	100	44	25
Patentscope	24	90	90	55	24	100	59	20
Google Patents	21	0	70	100	100	100	76	50
Google Scholar	17	0	50	75	64	60	50	45
Summon ČVUT	5	0	0	100	100	60	69	60
Scopus	60	80	100	70	100	80	46	50
Web of Science	24	80	100	100	95	80	35	60
Google News	34	10	50	40	100	0	–	–
Bing News	13	90	100	40	–	0	–	–
Technický týdeník	18	0	50	10	1	20	40	15

Kapitola 6

Určení vah metrik

Kapitola 4 identifikuje osm vyčíslitelných metrik a kapitola 5 hodnotí zdroje na základě těchto metrik. Lze ale předpokládat, že všechny metriky nebudou při vybírání zdroje pro forecasting nových technologií stejně významné. Například lze odhadnout, že ochota firem spolupracovat nemusí být do stejné míry důležitá jako samotný objem dat. Proto je nutné zavést váhy důležitosti nebo významnosti a přiřadit je každé metrice, čili metriky budou vážené. Tato kapitola se věnuje výpočtu a určení vah a implementaci výpočetního modelu na základě čehož bude určena kvalita zkoumaných zdrojů.

Cílem této kapitoly je získat váhy parametrů, vybrat model a pro získání konkrétních výsledků jej i implementovat. Na závěr bude možné spočítat a implementovat model pro výpočet samotné kvality konkrétních zdrojů, váhy metrik i jejich hodnoty potřebné pro výpočet budou k dispozici.

6.1 Výpočet vah

Jak je naznačeno v úvodu kapitoly, lze předpokládat, že všechny metriky nebudou pro výběr zdroje pro automatizovaný forecasting nových technologií stejně důležité. Je ale nutné rozhodnout o výběru toho správného.

Rozhodování je doménou projektových manažerů a je pro něj nezbytné vzít v potaz jednotlivé faktory a ty ohodnotit dle důležitosti. Při rozhodování vypomůže teorie vícekriteriálního rozhodování (Multiple-criteria decision-making). To si klade za cíl vybrat jednu z více zvažovaných variant (v tomto případě jsou varianty zdroje). Kritérii jsou v případě této práce kvantifikovatelné metriky. Po přiřazení vah jednotlivým kritériím lze vypočítat a vybrat nejlepší možnost, variantu. A právě přiřazením vah se vícekriteriální rozhodování také zabývá a právě z této z této teorie bude čerpáno.

Analýza a užití vícekriteriálního rozhodování v této práci čerpá z knihy „Manažerské rozhodovaní“ [42] a zejména z knihy „Rozhodování v managementu“ [43].

6.1.1 Výpočet na základě dotazníku

Původním plánem, jak zjistit váhy metrik, bylo použít lineární regresi popsanou dále a dotazníkové šetření mezi experty. Dotazníkové šetření by zjistilo odhady kvalit zdrojů pro vybrané zdroje a dle toho by lineární regrese stanovila váhy jednotlivých metrik.

Lineární regrese je model, který slouží k odhadu lineární závislosti jedné nebo více nezávislých proměnných na závislých proměnných. Po vložení nezávislých proměnných lze odhadnout závislou. Nezávislé jsou v případě práce metriky, z nichž každá má svůj regresní koeficient (váhu). Vzorec pro popsaný vztah vypadá následovně:

$$KvalitaZdroje = \beta_0 + \beta_1 * Trvanlivost + \beta_2 * OchSpol.. + \beta_7 * Relevance \quad (6.1)$$

Inspiraci pro strukturu a náležitosti dotazníku poskytla kniha „Survey Methodology“ [44]. Kniha poskytuje mnoho informací z oboru průzkumů a dotazování, pro potřeby této práce sice zachází až do přílišných detailů, nicméně je díky ní možné získat dobrý přehled o tématu. Vytvořený dotazník je k nahlédnutí v příloze A.

Ukázalo se ale, že lidé v okruhu autora (ze školy, z práce, známí), kteří jako respondenti přicházeli v úvahu, nepoužili často více než tři zdroje z dotazníku, navíc to byly většinou ty stejné zdroje napříč všemi respondenty. Předpokládaný počet respondentů byl původně 10. Nicméně po zjištění, že respondenti se zdroji nemají zkušenosti, bylo hledání dostatečného počtu respondentů s dostatečnými zkušenostmi vyhodnoceno jako velmi obtížné a v rozumné časové dotaci až neproveditelné.

Z toho důvodu bylo přikročeno k jinému způsobu určení vah metrik, které je popsané v následující sekci.

6.1.2 Metody výpočtu vah

Obecně se váhy stanovují na základě expertních odhadů. Jako vstupy se často pro určení vah užívají názory expertů získané například dotazníkovým šetřením. V případě této práce by byli potřeba experti pro kvalitu dat, kteří znají identifikované zdroje a jsou seznámeni s problémem forecastingu nových technologií. Po aktivním hledání autora jich nebylo nalezeno dostatečné množství k získání relevantních vstupů, proto bude k determinaci vah použita znalost autora. Popsané postupy jsou ale opakovatelné a znovupoužitelné i pro větší množství názorů.

Následují některé metody užívané pro zavedení vah při vícekriteriálním rozhodování. Tento obor je rozvinutý a zahrnuje mnoho dalších zde nezmíněných metod a jejich kombinací. Pro ilustraci jich bylo vybráno několik, z toho bude jedna použita pro demonstraci výpočtu vah.

Nejprve budou definovány podmínky, které musí váhy splňovat:

$$\sum_{i=0}^{n} w_i = 1 \tag{6.2}$$

$$0 \leq w_i \leq 1 \tag{6.3}$$

tedy váhy nabývají hodnot v rozmezí od 0 do 1 a jejich součet je 1, kdy n je počet vah a w_i je konkrétní váha pro kritérium (metriku).

Následují vybrané metody, obecně jsou vzorce určeny pro více expertů, následující jsou upraveny pouze pro výpočet pro jednoho hodnotícího.

6.1.2.1 Metoda pořadí

Jestliže je n počet kritérií, nejdůležitějšímu kritériu je dle experta přiřazena hodnota n, druhému nejdůležitějšímu $n-1$, dalšímu $n-2$ a tak dále. Přiřazené číslo se nazývá v. Váha důležitosti kritéria i je potom dána následujícím vztahem:

$$w_i = \frac{v_i}{\sum_{j=1}^{n} v_j} \tag{6.4}$$

6.1.2.2 Metoda bodování

Expert na základě zvolené bodové stupnice ohodnotí jednotlivá kritéria. Vyšší hodnota je přiřazena důležitějším kritériím, stupnice může být například od 0 do 10. Proměnná z reprezentuje bodové ohodnocení kritéria. Váha se vypočítá následujícím vztahem:

$$w_i = \frac{z_i}{\sum_{j=1}^{n} z_j} \tag{6.5}$$

6.1.2.3 Metoda párového porovnání

Tato metoda je vhodná i pro větší počet kritérií a lze ji provést strojově. Každé kritérium uvedené v r-tém řádku se srovnává s každým kritériem uvedeném v k-tém sloupci (mimo samo sebe, tedy když k=r).

Pokud hodnotitel považuje kritérium v r-tém řádku za důležitější než to v k-tém sloupci, zapíše do průsečíku r-tého řádku a k-tého sloupce hodnotu 1, v opačném případě 0. Součtem hodnot v řádku vznikne číslo u udávající důležitosti kritéria. Vlastní váha se vypočte dle následujícího vztahu:

$$w_i = \frac{u_i}{\sum_{j=1}^{n} u_j} \tag{6.6}$$

6.1.3 Aspirační úrovně

Aspirační úroveň je hodnota, které musí metrika dosáhnout, aby byl zdroj hodnocen jako potenciálně vhodný pro zamýšlené užití. Pokud této hodnoty metrika dosáhne, tak má smysl pro zdroj kvalitu počítat. V opačném případě je zdroj diskvalifikován a není hodnocen. Aspirační úrovně jsou minimální požadavky kladené na zdroje, které musí být splněny.

6.1.4 Výpočet

Pro demonstraci je vybrána metoda bodování. Metoda párového porovnání je vhodná spíše pro větší počet kritérií (více než 20) a metoda pořadí neumožňuje ohodnotit dvě kritéria stejnými body, což by mohlo být zavádějící.

Pro ohodnocení je zavedena desetibodová stupnice, kde vyšší hodnoty znamenají důležitější metriky. Také budou vyhodnoceny z pohledu aspiračních úrovní. Výsledné hodnoty se nachází v tabulce 6.1.

Trvanlivost

Trvanlivost je důležitá zejména z pohledu dlouhodobosti implementovaného řešení a zrušení funkčnosti zdroje znamená velkou změnu pro celý systém pro forecasting. Trvanlivost je ohodnocena osmi body z deseti.

Technické zpracování

Ochota spolupracovat je rozdílová metrika zejména v případě, že nabývá nulové hodnoty, potom zdroje nelze vůbec použít. Pokud poskytuje data nepraktickým způsobem, může prodražit implementaci, z toho důvodu je ohodnocena pouze pěti body.

Strojové zpracování je podobně důležité jako ochota spolupracovat, pouze však koriguje cenu, za kterou bude zdroj integrován do systému pro predikci. Nemá tak výrazný vliv na cenu jako Ochota spolupracovat, je ohodnocena pouze čtyřmi body.

Možnosti vyhledávání jsou podobně důležité jako Strojové zpracování, znovu ale ovlivní cenu implementace. Zásadním problémem je, pokud není možné vyhledat dokumenty zpět v čase, tím se stává zdroj nepoužitelný. Pokud tedy není splněna tato podmínka, měl by být zdroj diskvalifikován. Aspirační úroveň je možnost vyhledání dokumentů zpět v čase pro více než deset let (tato hodnota je diskutována v předchozích kapitolách). V opačném případě se jedná o metriku důležitou pouze z hlediska ceny implementace.

Hodnota informací

Objem dat je důležitým faktorem pro úspěšnou predikci. Lze předpokládat, že čím více dat bude ve zdroji k dispozici, tím více nových nápadů bude

možné nalézt. Tato metrika je ohodnocena sedmi body, jelikož ve vztahu k Trvanlivosti je méně důležitá. Je naopak důležitější oproti všem metrikám z Technického zpracování.

Reputace a Úroveň dat jsou stejně důležité jako Objem dat. Zejména proto, že se doplňují – čím více je k dispozici dat, tím horší by měly mít reputaci a obráceně. Čím lepší mají reputaci a jsou přísněji vybírané, tím méně by jich mělo být. Který zdroj nalezne rozumnou balanci mezi těmito dvěma metrikami by měl být kvalitní.

Potenciál zdroje pro predikci je nejdůležitější metrika, jelikož přímo a prokazatelně ovlivňuje schopnost zdroje predikovat nové technologie či myšlenky, a to je právě smyslem forecastingu nových technologií. Je to nejdůležitější metrika, která má největší dopad na predikce, je proto hodnocena deseti body.

Relevance vyhledaných položek k dotazu je podobně důležitá metrika jako Potenciál zdroje pro predikci, jelikož umožňuje nalézt dostatečně vhodné výsledky, ze kterých vznikají predikce. Na druhou stranu zkoumá hlavně několik prvních nalezených dokumentů, tedy ověřuje funkčnost vyhledávače. Je možné, že se relevantní dokumenty budou nacházet i na dalších pozicích, i když by správně neměly, nejrelevantnější výsledky by měly být na prvních pozicích ve vyhledávači. Metrika je hodnocena osmi body.

Název metriky	z_i	w_i
Trvanlivost	8	0,154
Ochota spolupracovat	5	0,096
Strojové zpracování	4	0,077
Možnosti vyhledávání	3	0,058
Objem dat	7	0,135
Reputace, Úroveň dat	7	0,135
Potenciál zdroje pro predikci	10	0,192
Relevance vyhledaných položek k dotazu	8	0,154

Tabulka 6.1: Metriky ohodnocené body důležitosti a příslušné váhy

Hodnoty w_i jsou zaokrouhleny na tři desetinná místa. Vznikly aplikováním vzorce 6.5 na bodové ohodnocení metrik.

Celkem existuje jedna aspirační úroveň, pro Možnosti vyhledávání je to možnost vyhledat dokumenty minimálně deset let zpět v čase.

6.2 Výpočet kvality zdrojů

Nyní jsou k dispozici všechny vstupy pro výpočet kvality zdrojů. Vícekriteriální rozhodování znovu nabízí několik cest, kterými se lze při výběru varianty (zdroje) vydat. Jsou popsány tři obecné metody a jedna metoda uzpůsobená přímo zavedenému hodnocení metrik z této práce. Existují i další metody, ale

nejsou popsané například proto, že nejsou vhodné pro vyšší počet kritérií (více než 5) nebo jsou například jen mírnou úpravou popsaných metod.

6.2.1 Metody výpočtu kvality

Na následujících řádcích již bude terminologie přizpůsobená úloze řešené v této práci. V sekci 5.7 na konci kapitoly 5 je k dispozici přehled všech hodnot metrik pro vyčíslené zdroje.

6.2.1.1 Metoda pořadí

Jak již název napovídá, stěžejním nástrojem této metody budou operace s pořadím. Každá metrika se oboduje pořadím a to se vynásobí váhou. Čísla na řádku se potom sečtou a vzniknou tak hodnoty kvality zdrojů, kde nižší číslo je lepší. Pro vyčíslené zdroje by implementace této metody vypadala následovně.

Dle aspiračních kritérií definovaných v předchozí sekci se nekvalifikovaly Google News a Bing News, proto nebudou při vyhodnocení brány v úvahu.

Následující tabulka 6.2 obsahuje hodnocení vyčíslených zdrojů a pořadí jejich metrik. Názvy metrik jsou zkrácené, kompletní tabulka je k nalezení v sekci 5.7. Sloupec VSP je vážený součet pořadí, tedy kvalita zdroje, a sloupec nejvíce vpravo s názvem „p“ značí pořadí zdrojů dle VSP.

Název zdroje	**T**	**OS**	**SZ**	**MV**	**OD**	**R**	**P**	**R**	**VSP**	**p**
Espacenet	2	1	1	4	4	1	6	4	2,864	5
Patentscope	3	2	2	4	5	1	3	5	2,845	4
Google Patents	4	4	3	1	1	1	1	2	1,783	1
Google Scholar	6	4	4	2	3	3	4	3	3,384	7
Summon ČVUT	7	4	5	1	1	3	2	1	2,562	3
Scopus	1	3	1	3	1	2	5	2	2,128	2
Web of Science	3	4	1	1	2	2	8	1	3,080	6
Technický týd.	5	4	4	5	6	4	7	6	4,861	8

Tabulka 6.2: Souhrnná tabulka pořadí metrik a kvalit dle metody pořadí

6.2.1.2 Bodovací metoda

Toto je jiná bodovací metoda než použitá v předchozí sekci. Tato metoda pracuje s definicí jednotlivých intervalů pro každou metriku. Pro každou metriku by bylo stanoveno několik rozsahů hodnot, které by byly obodovány. Tato metoda se hodí pro zjednodušení rozsahů hodnot, které nejsou standardizované. Pro standardizované metriky ale nemá smysl zavádět rozsahy hodnot a ty bodovat.

6.2.1.3 Metoda váženého součtu

Každé hodnotě metriky bude přiřazena užitková funkce u_i, která bude nabývat hodnot z intervalu $\langle 0; 1\rangle$ pro metriku i. Tato metoda je především vhodná pro kvantitativní kritéria a tudíž i pro metriky z této práce. Hodnotou 1 bude ohodnocena nejvyšší hodnota metriky a naopak hodnotou 0 nejnižší. Hodnoty metrik budiž značeny y_i, užitková funkce pak bude:

$$u_i = \frac{y_i - d}{h - d} \tag{6.7}$$

Kdy h je nejvyšší hodnota, kterou metrika nabývá, a d je nejnižší. Pokud bude n počet metrik, pak pro kvalitu zdroje (kz) platí, že:

$$kz = \sum_{i=1}^{n} (w_i * u_i) \tag{6.8}$$

Postup je demonstrovaný v následující tabulce 6.3, kdy KZ je kvalita zdroje a sloupec nejvíce vpravo s názvem „p“ značí pořadí zdrojů dle KZ, čísla jsou oříznuta na dvě desetinná místa.

Název zdroje	**T**	**OS**	**SZ**	**MV**	**OD**	**R**	**P**	**R**	**KZ**	**p**
Espacenet	0,81	1	1	0,5	0,59	1	0,21	0,22	0,61	4
Patentscope	0,34	0,9	0,9	0,5	0,23	1	0,58	0,11	0,53	6
Google Patents	0,29	0	0,7	1	1	1	1	0,77	0,74	2
Google Scholar	0,21	0	0,5	0,72	0,63	0,5	0,36	0,66	0,44	7
Summon ČVUT	0	0	0	1	1	0,5	0,82	1	0,57	5
Scopus	1	0,8	1	0,66	1	0,75	0,26	0,77	0,75	1
Web of Science	0,34	0,8	1	1	0,94	0,75	0	1	0,64	3
Technický týd.	0,23	0	0,5	0	0	0	0,12	0	0,09	8

Tabulka 6.3: Souhrnná tabulka pořadí metrik a kvalit dle metody váženého součtu

U každé metriky byly nalezeny největší a nejnižší hodnoty, a na základě vzorce 6.7 byly vypočítány užitky jednotlivých metrik pro zdroje. Ty jsou obsaženy v tabulce výše. Byly vynásobeny váhami a na základě tohoto výpočtu vznikla hodnota Kvality Zdroje (KZ).

6.2.1.4 Metoda váženého součtu bez další normalizace

Tato metoda je vlastní variace na předchozí metodu váženého součtu. Jelikož hodnoty jsou již normalizované na stupnici od 0 do 100, nemusí být nutné je normalizovat znovu a zavádět novou nejnižší a novou nejvyšší hodnotu. Následující tabulka 6.4 ukazuje pořadí zdrojů, kde jsou místo užitků reálně využity naměřené hodnoty metrik. Čím vyšší číslo, tím vyšší má zdroj kvalitu zdroj.

Název zdroje	T	OS	SZ	MV	OD	R	P	R	KZ	p
Espacenet	50	100	100	55	60	100	44	25	75,76	4
Patentscope	24	90	90	55	24	100	59	20	65,59	5
Google Patents	21	0	70	100	100	100	76	50	83,79	3
Google Scholar	17	0	50	75	64	60	50	45	52,63	6
Summon ČVUT	5	0	0	100	100	60	69	60	44,85	7
Scopus	60	80	100	70	100	80	46	50	89,01	2
Web of Science	24	80	100	100	95	80	35	60	95,62	1
Technický týd.	18	0	50	10	1	20	40	15	17,83	8

Tabulka 6.4: Souhrnná tabulka pořadí metrik a kvalit dle metody váženého součtu bez další normalizace

6.3 Diskuze výsledků

V metodě pořadí se smazávají velké rozdíly mezi dvěma zdroji, pokud by byly porovnávány dva zdroje, jeden s deseti miliony dokumenty a druhý s tisícem dokumentů, rozdíl by mezi nimi byl pouhý jeden bod v pořadí. Větší výpovědní hodnotu mají zbylé dva vyčíslené způsoby.

Následující tabulka 6.5 shrnuje pořadí dle všech tří vyčíslených metod a stanovuje finální pořadí zdrojů. Pořadí jsou jednoduše sečtena, čím je pořadí nižší, tím má zdroj vyšší kvalitu. Celkové pořadí zdrojů uvádí finální pořadí zdrojů dle kvality.

Název zdroje	Metoda pořadí	Metoda váženého součtu	Metoda váženého součtu bez další normalizace	Součet pořadí	Celkové pořadí zdrojů
Espacenet	5	4	4	13	4
Patentscope	4	6	5	15	5
Google Patents	1	2	3	6	2
Google Scholar	7	7	6	20	7
Summon ČVUT	3	5	7	15	6
Scopus	2	1	2	5	1
Web of Science	6	3	1	10	3
Google News	–	–	–	–	–
Bing News	–	–	–	–	–
Technický týdeník	8	8	8	24	8

Tabulka 6.5: Vypočtená kvalita zdrojů

Google News a Bing News nesplnily aspirační úroveň pro vyhledávání zpět v čase, nejsou tedy hodnoceny. Jak je popsáno v kapitole 5, autor kontaktoval podporu vyhledávače Summon ČVUT s dotazem, jakou poskytuje službu pro strojové zpracování a jaká je jejich ochota spolupracovat. Bohužel i přes to, že odpovědět bylo možno déle než dva týdny, do doby dokončení této práce se autorovi odpovědi nedostalo. Proto je také tento vyhledávač hodnocen z pohledu metriky Ochota spolupracovat špatně.

Dle navržených metrik a provedeného hodnocení je pro úlohu automatizovaného forecastingu nových technologií v identifikované konkurenci nejlepší databáze Scopus od společnosti Elsevier.

Jako nejhorší zdroj je kromě těch, které se ani nedostaly do finálního hodnocení, hodnocen Technický týdeník, zejména kvůli špatnému vyhledávání a malému objemu dokumentů, což ztěžuje kvalitní predikci. Souhrnné pořadí je pro přehlednost prezentováno v následující tabulce 6.6, kde jsou zdroje seřazeny dle kvalit vypočtených na základě návrhu z této práce od nejlepšího (Scopus) po nejhorší (Technický týdeník a diskvalifikované zdroje).

Název zdroje	**Celkové pořadí**
Scopus	1
Google Patents	2
Web of Science	3
Espacenet	4
Patentscope	5
Summon ČVUT	6
Google Scholar	7
Technický týdeník	8
Google News	–
Bing News	–

Tabulka 6.6: Seřazené zdroje dle vypočtené kvality

Závěr

První část práce analyticky pokrývá forecasting a jeho varianty, systémy pro automatizovaný forecasting nových technologií, forecasting z ekonomického pohledu a měření kvality dat v různých oborech.

Podstatou této práce je vytvoření obecného a opakovatelného rámce pro výběr zdroje pro automatizovaný forecasting nových technologií, který je znovupoužitelný. Tento rámec je tvořen metrikami, váhami a metodami výpočtu.

Stěžejní částí rámce je návrh originálních metrik pro měření kvality zdrojů. Tento návrh vychází z analýzy v první části práce. Navržené metriky byly aplikovány na zdroje a tím byla ověřena jejich funkčnost a uplatnitelnost. Metriky byly na základě aplikace také upraveny do stávající podoby. Pro kompletnost celé metodiky měření kvality zdrojů byly metriky ohodnoceny váhami a ukázkově vypočítány kvality zkoumaných zdrojů.

Metriky byly navržené s důrazem na objektivitu a znovupoužitelnost. Všechny provedené pokusy a měření v této práci jsou opakovatelné. Kvalita zdrojů pro automatizovaný forecasting nových technologií nebyl přesně definovaný pojem, tato práce představuje vlastní rámec, jak kvalitu těchto zdrojů vyčíslit. Tato práce předkládá opodstatněné, vysvětlené a ukázkové užití definovaného rámce.

Navazující výzkum

Způsobů, jak na tuto práci navázat, je několik. První možností je podrobit výzkumu více zdrojů, než je v této práci identifikováno, a otestovat jejich chování vůči definovaným metrikám.

Druhou možností je držet se rámce a vyzkoušet jiné modely pro výpočty kvality. Všechny metodiky pro získání vstupů jsou v této práci definované.

Třetí možností je počítat váhy metrik jinými než popsanými způsoby, například pomocí pokročilých technik jako jsou metody umělé inteligence – k-nejbližších sousedů (k-nearest neighbors), Bayesovský klasifikátor (Bayes classifier), a to za pomoci názorů více expertů.

Dosažené výsledky

Tato práce také otevírá nejen pro FIT ČVUT zajímavé téma predikce nových technologií.

Výsledky konkrétních výpočtů jsou diskutovány na konci kapitoly 6 v sekci 6.3. Kromě těchto výsledků se podařilo definovat novou metodiku pro určování kvality zdrojů pro automatizovaný forecasting nových technologií.

Práce obsahuje i objemné analýzy, kromě zevrubné analýzy samotného forecastingu i detailní analýzu nástroje pro automatizovaný forecasting nových technologií. Jeho vývoj je popsaný ve více než deseti odborných článcích, zde je vývoj shrnutý. Může to být cenný zdroj informací pro jedince začínající s tímto tématem.

Analýza ekonomického pohledu je kromě popisů přínosů forecastingu pro podnikovou sféru zejména členěním firem zabývajících se forecastingem (2.2.1). Toto členění je originální a vzniklo na základě důkladného průzkumu firem provozujících forecasting a ekonomických dopadů forecastingu.

Zejména ale práce obsahuje originální metodiku s opakovatelnými výpočty a testy. Celkem je tedy k dispozici přesný vzorec a definice metrik, dle čehož lze vypočítat kvalitu libovolného zkoumaného zdroje.

Dopady výsledků práce na ekonomický přínos

Jak je popsáno v kapitole 1 a 2, výběr správného zdroje je pro automatizovaný forecasting nových technologií jednou z důležitých částí úspěchu podniku. Vhodně zvolený zdroj zjednoduší implementaci nástroje, jelikož nemusí být potřeba složité integrace nebo mnoho funkcí potřebných pro predikci poskytuje v základu sám. Dále je trvanlivý a díky tomu se zmenšuje šance na jeho znepřístupnění (jak se stalo v případu zkoumanému v kapitole 1). Dobře zvolený zdroj také zlepšuje funkce samotného forecastingu, jelikož obsahuje relevantní data, na základě kterých je možné provádět kvalitní predikce.

Správně vybraný datový zdroj zlevní vývoj, zlepší funkci samotného systému pro automatizovaný forecasting nových technologií a ušetří projektové zdroje na údržbu celého systému.

Tato práce poskytuje pro výběr toho správného zdroje metodiku. Autoři systémů pro forecasting se nad výběrem zdroje buď příliš nerozmýšleli nebo vyzkoušeli integraci více nástrojů a dle toho vybírali. To ale nepokryje všechny definované metriky a navíc je takový přístup velice nákladný. Tato práce tak zlevňuje a zefektivňuje výběr správného zdroje pro automatizovaný forecasting nových technologií.

Srovnání odvedené práce se zadáním

Tato sekce poskytuje porovnání zadání a odvedené práce.

1. Formulujte pojem forecastingu. Analyzujte možnosti automatizovaného forecastingu nových technologií.

Pojem forecastingu je formulován v kapitole 1, věnuje se členění forecastingu, samotné definici a úvodu do metod forecastingu. Od sekce 1.4 se kapitola věnuje forecastingu nových technologií.

2. Analyzujte ekonomické přínosy/dopady automatického forecastingu při strategickém plánování budoucího směřování podniku a zavádění nových technologií.

Kapitola 2 se věnuje ekonomickým aspektům forecastingu, kde v první čísti definuje nezbytné pojmy a zkoumá jeho dopady. Popisuje trh s forecastingem (2.2), zejména zavádí originální vlastní dělení firem zabývajících se forecastingem (2.2.1). Na závěr zkoumá jeho význam pro zavádění nových technologií v podniku (2.3).

3. Identifikujte a analyzujte kvantitativní zdroje potenciálně vhodné pro automatizovaný forecasting se zaměřením na predikci nových technologií.

Kapitola 3 definuje datový zdroj. Identifikuje a analyzuje celkem 16 datových zdrojů rozdělených do třech skupin.

4. Navrhněte metriky pro měření kvality kvantitativních zdrojů pro automatizovaný forecasting pro predikci nových technologií a aplikujte je na identifikované zdroje.

Kapitola 4 začíná seznámením s pojmem kvality dat (4.1), v této práci je to chápáno jako synonymum ke kvalitě zdroje. Po analýze možností měření kvality dat pro různé obory jsou navrženy originální vlastní metriky pro měření kvality zdrojů pro automatizovaný forecasting nových technologií (4.3).

V kapitole 5 jsou aplikovány na některé z identifikovaných zdrojů. Výběr konkrétních zdrojů je zdůvodněn na začátku kapitoly. Na základě aplikace na zdroje byly metriky upraveny do aktuální podoby, tento postup je popsán na konci kapitoly (5.6).

5. Zvolte výpočetní model pro ohodnocení navržených metrik váhami, implementujte jej a stanovte příslušné váhy.

V kapitole 6 byl zvolen a implementován model pro výpočet vah a váhy byly přiřazeny jednotlivým metrikám. Sekce 6.2.1 obsahuje vypočteny kvality pro vyčíslené zdroje.

6. Diskutujte dosažené výsledky, zejména možné dopady na ekonomický přínos.

Dosažené výsledky jsou diskutovány na konci kapitoly 6 (6.3) a zde v Závěru práce.

Literatura

[1] Madnick, S.; Woon, W. L.; Firat, A.: Technological Forecasting - A Review. 2008, [online, cit. 6.3.2018]. Dostupné z: http://web.mit.edu/smadnick/www/wp/2008-15.pdf

[2] Investopedia: [online, cit. 6.3.2018]. Dostupné z: https://www.investopedia.com/terms/f/forecasting.asp

[3] Webster, M.: Dictionary. Encyclopædia Britannica Inc., [online, cit. 24.2.2018]. Dostupné z: https://www.merriam-webster.com/dictionary/forecast

[4] Abraham, B.; Ledolter, J.: *Statistical Methods for Forecasting*. Wiley, 2009.

[5] Derby, N.: Time Series Forecasting Methods. In *Statis Pro Data Analytics*, Calgary SAS Users Group, December 2009 [online, cit. 20.2.2018]. Dostupné z: https://www.sas.com/content/dam/SAS/en_ca/User%20Group%20Presentations/Calgary-User-Group/Derby-TimeSeriesForecasting-Nov2009.pdf

[6] IBM: IBM Watson Health. International Business Machines Corporation, [online, cit. 20.2.2018]. Dostupné z: https://www.ibm.com/watson/health/

[7] Ziegler, B.: Methods for Bibliometric Analysis of Research: Renewable Energy Case Study. 2009, [online, cit. 25.3.2018]. Dostupné z: http://web.mit.edu/smadnick/www/wp/2009-10.pdf

[8] Madani, F.; Weber, C.: The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis. 2016, [online, cit. 24.2.2018]. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0172219016300412

[9] Porter, A. L.; Cunningham, S. W.; Banks, J.; aj.: *Forecasting and Management of Technology, 2nd Edition*. Wiley, 2011.

[10] Lupu, M.; Mayer, K.; Kando, N.; aj.: *Current Challenges in Patent Information Retrieval*. Springer-Verlag GmbH Germany, 2011.

[11] Madnick, S.; Woon, W. L.: Research plan for semantics-enabled technology forecasting: A case study on alternative energies. 2007, [online, cit. 6.3.2018]. Dostupné z: web.mit.edu/smadnick/www/wp/2007-14.pdf

[12] Madnick, S.; Woon, W. L.; Henschel, A.; aj.: Technology Forecasting Using Data Mining and Semantics: Second Annual Report. 2009, [online, cit. 6.3.2018]. Dostupné z: web.mit.edu/smadnick/www/wp/2009-15.pdf

[13] Madnick, S.; Woon, W. L.: Semantic Distances for Technology Landscape Visualization. 2008, [online, cit. 18.3.2018]. Dostupné z: http://web.mit.edu/smadnick/www/wp/2008-04.pdf

[14] Kyebambe, M. N.; Cheng, G.; Huang, Y.; aj.: Forecasting emerging technologies: A supervised learning approach through patent analysis. Technological Forecasting and Social Change, 2017, [online, cit. 15.4.2018]. Dostupné z: https://www.deepdyve.com/lp/elsevier/forecasting-emerging-technologies-a-supervised-learning-approach-0L0LXdLl3Y?articleList=%2Fsearch%3Fquery%3Dforecasting%2Bemerging%2Btechnologies

[15] Johnson, G.; Scholes, K.; Whittington, R.: *Exploring Corporate Strategy: Text & Cases*. Prentice-Hall, 2004.

[16] Pavlickova, P.: Strategicke rizeni informatiky MI-SMI. Czech Technical University, Faculty of Information Technology, MI-SMI, lecture 8., 2017.

[17] Pavlickova, P.: Strategicke rizeni informatiky MI-SMI. Czech Technical University, Faculty of Information Technology, MI-SMI, lecture 6., 2017.

[18] International Organization for Standardization: ISO 31000 - Risk managementI. [online, cit. 5.3.2018]. Dostupné z: https://www.iso.org/iso-31000-risk-management.html

[19] Efendigil, T.; Onut, S.; Kahraman, C.: A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis. 2008, [online, cit. 10.3.2018]. Dostupné z: https://pdfs.semanticscholar.org/33dc/1b386f0a79be6bd7b4bea7e7aadd222f98ad.pdf

[20] Aulkemeier, F.; Daukuls, R.; Iacob, M.-E.; aj.: Sales Forecasting as a Service. 2016, [online, cit. 10.3.2018].

Dostupné z: https://pdfs.semanticscholar.org/8877/7896e308b819b28dd96ca04d55f02af9d6f3.pdf

[21] Gartner Inc.: Hype Cycle. 2008, [online, cit. 10.3.2018]. Dostupné z: https://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

[22] Fenn, J.: 2010 Emerging Technologies Hype Cycle is Here. Gartner Inc., 2010, [online, cit. 30.3.2018]. Dostupné z: https://blogs.gartner.com/hypecyclebook/2010/09/07/2010-emerging-technologies-hype-cycle-is-here/

[23] Gartner's 2016 Hype Cycle for Emerging Technologies Identifies Three Key Trends That Organizations Must Track to Gain Competitive Advantage. Gartner Inc., 2016, [online, cit. 22.4.2018]. Dostupné z: https://www.gartner.com/newsroom/id/3412017

[24] Herschel, G.; Goldman, M.: Getting 'Gartnered': How Vendors Can Work With Gartner. 2008, [online, cit. 10.3.2018]. Dostupné z: https://www.gartner.com/it/about/max_time/Getting_Gartnered_How_Vendors_Can_Work_With_Gartner.pdf

[25] Porter, A. L.; Cunningham, S. W.: *Tech Mining: Exploiting New Technologies for Competitive Advantage*. Wiley, 2004.

[26] Fontana, R.; Nuvolari, A.; Shimizu, H.; aj.: Reassessing patent propensity: evidence from a data-set of R&D awards 1977-2004. 2013, [online, cit. 19.3.2018]. Dostupné z: http://ideas.repec.org/p/ise/isegwp/wp092013.html

[27] Orduna-Malea, E.; Ayllon, J. M.; Martin-Martin, A.; aj.: About the size of Google Scholar: playing the numbers. 2014, [online, cit. 19.3.2018]. Dostupné z: https://arxiv.org/abs/1407.6239

[28] Beel, J.; Gipp, B.: Academic Search Engine Spam and Google Scholars Resilience Against it. 2010, [online, cit. 25.3.2018]. Dostupné z: https://quod.lib.umich.edu/j/jep/3336451.0013.305?rgn=main;view=fulltext

[29] Another fake paper has been accepted for publication and oral presentation. 2009, [online, cit. 25.3.2018]. Dostupné z: http://diehimmelistschoen.blogspot.cz/2009/01/another-paper-has-been-accepted-for.html

[30] Olson, J. E.: *Data Quality: The Accuracy Dimension*. Elsevier, 2003.

[31] TOLE, A. A.: Big Data Challenges. 2013, [online, cit. 25.2.2018]. Dostupné z: http://www.dbjournal.ro/archive/13/13_4.pdf

[32] Gandomi, A.; Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. 2015, [online, cit. 26.3.2018]. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0268401214001066

[33] Cai, L.; Zhu, Y.: The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. 2015, [online, cit. 27.3.2018]. Dostupné z: https://datascience.codata.org/articles/10.5334/dsj-2015-002/

[34] Hox, J. J.; Boeije, H. R.: Data Collection, Primary vs. Secondary. 2005, [online, cit. 26.3.2018]. Dostupné z: https://dspace.library.uu.nl/bitstream/handle/1874/23634/hox_05_data+collection,primary+versus+secondary.pdf?sequence=1

[35] van Nederpelt, P. W. M.; Daas, P. J. H.: 49 Factors that influence the quality of secondary data sources. 2012, [online, cit. 26.3.2018]. Dostupné z: https://unstats.un.org/unsd/dnss/docs-nqaf/49%20Factors%20that%20influence%20the%20Quality%20of%20Secondary%20Data%20Sources.pdf

[36] World Wide Web Consortium: Linked Data. [online, cit. 27.3.2018]. Dostupné z: https://www.w3.org/standards/semanticweb/data

[37] Bizer, C.; Heath, T.; Berners-Lee, T.: Linked Data - The Story So Far. 2009, [online, cit. 27.3.2018]. Dostupné z: http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

[38] Zaveri, A.; Rula, A.; Maurino, A.; aj.: Quality Assessment for Linked Data: A Survey. 2012, [online, cit. 27.3.2018]. Dostupné z: http://www.semantic-web-journal.net/system/files/swj773.pdf

[39] Lee, Y. W.; Strong, D. M.; Kahn, B. K.; aj.: AIMQ: a methodology for information quality assessment. 2001, [online, cit. 27.3.2018]. Dostupné z: http://web.mit.edu/tdqm/www/winter/AIMQ.pdf

[40] Daepp, M. I. G.; Hamilton, M. J.; West, G. B.; aj.: The mortality of companies. 2015, [online, cit. 30.3.2018]. Dostupné z: http://rsif.royalsocietypublishing.org/content/12/106/20150120

[41] Noorden, R. V.: Global scientific output doubles every nine years. 2014, [online, cit. 29.3.2018]. Dostupné z: blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html

[42] Fotr, J.; Dedina, J.: *Manazerske rozhodovani*. VSE Praha, 1993.

[43] Zacek, V.: *Rozhodovani v managementu*. CVUT Praha, 2015.

[44] Groves, R. M.: *Survey Methodology*. Wiley, 2004.

Příloha **A**

Dotazník

Dotazník – zdroje pro forecasting

Úvod

Tento dotazník slouží pro zjištění expertních odhadů.

DP se věnuje detekci nových technologií, o kterých veřejnost ještě neví. Zmínky o nich jsou právě pouze v patentech nebo pár odborných článcích.

Metoda forecastingu

Jak tedy předvídat? Podívat se do nějakého zdroje (databáze) na dokumenty (patenty, odborné články, zprávy), kde by se zmínky o technologiích mohly vyskytovat a hledat výrazy, které jsou nové a s přibývajícím časem se objevují čím dál častěji (je jim věnováno více článků) – tyto výrazy by měly být názvy nových technologií.

Dotazník

V práci hodnotím zdroje podle osmi metrik:

- Trvanlivost
- Ochota Spolupracovat
- Strojové Zpracování
- Možnosti Vyhledávání
- Objem Dat
- Reputace, Úroveň dat
- Potenciál zdroje pro predikci
- Relevance vyhledaných položek k dotazu

Vás prosím o expertní ohodnocení následujících zdrojů na škále od 0 do 100, podle Vašich znalostí. Do jaké míry jsou dle Vás následující zdroje pro popsanou úlohu vhodné?

0 – zcela nevhodný zdroj

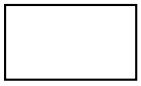
100 – nejvhodnější zdroj

Data o patentech

PatentScope

Patentscope databáze obsahuje 64 milionů patentových dokumentů včetně 3,1 milionů zveřejněných patentových přihlášek. Patentová přihláška je žádost o uznání patentu. To znamená, že v databázi jsou i patenty v procesu schvalování.

Patentscope je vlastněn WIPO - World Intellectual Property Organization, což je organizace sdružující skoro všechny země světa (189 zemí).

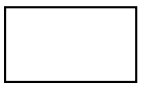


https://patentscope.wipo.int

Espacenet

Espacenet je databáze vlastněná EPO - Evropským patentovým úřadem. Jeho celosvětová databáze obsahuje přes 100 mil. patentových dokumentů, což je ze všech zmíněných zdrojů patentů nejvíce.

Espacenet nabízí přístup k informacím o vynálezech a patentech datovaných od 19. století dodnes. Součástí jsou i české patenty.



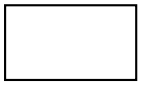
https://worldwide.espacenet.com

Google Patents

Google Patents indexuje patenty jak od již zmíněných EPO (Evropský patentový úřad) a WIPO (World Intellectual Property Organization), tak i od dalších patentových organizací. Mezi ně patří asijské úřady (Čína, Japonsko, Korea), jejichž patenty byly přeloženy do angličtiny a jsou tak lépe vyhledatelné.

Dle oficiálních informací dokáže vyhledávat mezi více než 87 miliony patenty. Vyvstává otázka, proč je toto číslo nižší než u Espacenet. Je to proto, že Espacenet vyhledává ve většině evropských patentů, Google Patents ale pouze v patentech nejprodukti vnějších zemí Evropy, co do počtu patentů.

Služba Google Patents byla spuštěna v prosinci roku 2006 a od té doby přibývá počet úřadů, jejichž patenty indexuje.



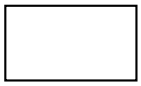
https://patents.google.com

Zdroje s odbornými znalostmi

Google Scholar

Google Scholar funguje podobně jako Google Patents na principu indexování dokumentů z jiných zdrojů. Indexuje dostupné odborné články, knihy, konferenční příspěvky, bakalářské, diplomové a disertační práce, technické zprávy, a další odbornou literaturu z různých zdrojů.

Google nezveřejňuje velikost obsahu, ale odborné odhady z roku 2014 hovoří o 160 miliónech záznamů.



https://scholar.google.cz

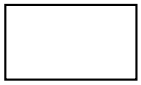
Vyhledávač Summon ČVUT

Vyhledávač má k dispozici 84 milionů výsledků, z toho 37 milionů článků v odborných časopisech, 28 milionů novinových článků a přes 230 000 knih. Je provozován Ústřední knihovnou ČVUT. Zajímavostí je, že obsahuje 30 milionů recenzovaných dokumentů, zbylé jsou například i novinové články.

www.cvut.summon.serialssolutions.com

Scopus

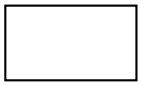
Databáze recenzované literatury vlastněná společností Elsevier, jež funguje jako vydavatelství i správce a archivátor dokumentů. Scopus obsahuje abstrakty a citace z vědeckých žurnálů, knih a konferencí, patenty a několik stovek obchodních publikací.



https://www.elsevier.com/solutions/scopus

Web of Science

Web of science je služba provozovaná korporací Thomson Reuters. Je přístupná po přihlášení přes partnerskou instituci (i ČVUT) nebo po registraci. Obsahuje 90 miliónů záznamů - obsahově stejných jako poskytuje Scopus.

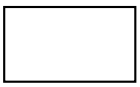


https://apps.webofknowledge.com

Zprávy

Google News

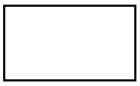
Google news prohledává přes 350 milionů článků a agreguje přes 4500 světových zdrojů. Časopisy, noviny, zpravodajství - na to vše se Google News zaměřuje.



https://news.google.cz

Bing News

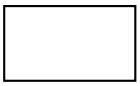
Konkurence Google News s mnohem menší základnou - zhruba 30 milionů článků. Původně známé také jako MSN News vlastněná Microsoftem.



www.bing.com/news

Technický týdeník

Český zástupce v přehledu zdrojů, s úctyhodnými šestnácti a půl tisíci technickými články může být zajímavou alternativou k zahraničním zdrojům.



http://www.technickytydenik.cz

Příloha **B**

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `src`
  - `thesis.tex` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF