Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Katedra informatiky a výpočetní techniky

Bakalářská práce

Predikce vítěze sportovního utkání využitím PageRanku

Plzeň, 2014 **Pavel Suda**

Prohlášení

Prohlašuji, že jsem Bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 23. června 2014

Pavel Suda

Poděkování

Na tomto místě bych rád poděkoval panu Ing. Michalu Nyklovi za vstřícnost a ochotu při vedení práce, cenné rady a čas, který mi byl věnován při konzultacích.

Abstract

Prediction of Sport Match Winner Using PageRank

Bachelor thesis is concerned with the prediction of the sport‘s match winner using the algorithm PageRank. In presented bachelor thesis you can find the description of the method for the predtiction of future edge in graph. In this thesis there is a brief description of the algorithm PageRand and following problems with its implementation. We can also find the description of sport competitions in which the prediction using PageRank was tested. Afterwards, there is a description of designed graphs and proposed modifications of personalized vector. This thesis contains the description of two applications which were created for the purpose of obtaining and presenting results of the prediction. At the end of the thesis, the success rate of obtained prediction results is discussed.

Predikce vítěze sportovního utkání využitím PageRanku

Práce se zabývá predikcí vítěze sportovního utkání využitím algoritmu PageRank. Naleznete zde popis používaných metod pro predikci budoucí hrany v grafu. V práci se nachází stručný popis algoritmu PageRank a problémů s jeho implementací. Jsou zde také popsány sportovní soutěže, na nichž je testována úspěšnost predikce využitím PageRanku. Dále se v práci nachází popis navržených typů grafů a navržené úpravy vektoru personalizace. Práce obsahuje popis dvou aplikací, které byly vytvořeny za účelem získání a prezentace získaných výsledků predikce. Na konci práce je diskutována úspěšnost získaných výsledků predikce.

Obsah

1 Úvod

Jedním z problémů při studii komplexních sítí, je predikce budoucí hrany. Výsledkem této práce by měla být aplikace, která bude predikovat vítěze jednotlivých sledovaných sportovních utkání, tedy tvořit nové hrany v síti (dále jen grafu). Predikce vítězů bude prováděna s využitím algoritmu PageRank, s možností výběru ze tří typů grafů a využití personalizace vrcholů. Jako referenční predikce bude sloužit predikce vítězů podle aktuálního postavení v tabulce.

V prvních kapitole jsou popsány metody pro predikci budoucí hrany v grafu. Po této úvodní kapitole, která nás seznámí se základy predikce budoucí hrany v grafu, následuje kapitola, která je věnována algoritmu PageRank. Kapitola obsahuje stručné vysvětlení problematiky implementace a použití algoritmu. Dále nás seznámí s problémy, které je nutné při implementaci vyřešit a nastíní nám jejich řešení.

Ve čtvrté kapitole se seznámíme s odvětvími sportů a s jejich konkrétními soutěžemi, na kterých bude testována úspěšnost predikce budoucí hrany. Soutěže jsou zde podrobně popsány a je zde uvedeno, co nás vedlo právě k tomuto výběru.

Pátá kapitola je věnována navrženým typům grafů, na nichž budou testovány a následně porovnávány výsledky. Následující kapitola je věnována samotné aplikaci pro predikci vítězů sledovaných sportovních utkání. Je zde podrobně popsán celý průběh aplikace, struktura vstupního a výstupního souboru doplněná o obrázky, pro lepší představu. Také zde najdeme popis predikce podle aktuálního postavení v tabulce, která je použita jako referenční.

Sedmá kapitola popisuje aplikaci pro vykreslování grafů ze získaných výsledků, která slouží pro lepší prezentaci získaných výsledků. Tato aplikace je v dané kapitole podrobně popsána.

V poslední kapitole je místo věnováno diskuzi získaných výsledků, kde jsou porovnávány a analyzovány získané výsledky. Také je zde diskutováno použití algoritmu PageRank v této problematice.

2 Metody predikce budoucí hrany v sociálních a jiných sítích

Mnoho sociálních, biologických a informačních systémů může být dobře popsáno grafem. Vrcholy grafu mohou reprezentovat jednotlivce, biologické prvky (geny), počítače, internetové uživatele atd. a hrany reprezentují vztahy mezi danými vrcholy. Studie komplexních grafů se proto staly jedním ze společných cílů mnoha odvětví věd a bylo vyvinuto velké úsilí k pochopení jejich vývoje [1,2]. Jedním z problémů je predikce hrany v grafu, což je snaha o odhadnutí pravděpodobnosti vytvoření nové hrany mezi dvěma vrcholy. Odhad predikce hrany je založen na zjištěných vazbách a atributech daných vrcholů [3].

Kromě využití při analýze grafů s chybějícími údaji, mohou být algoritmy pro predikci hrany použity k předpovědi hran, které se mohou objevit v budoucím vývoji grafu. Např. v on-line sociálních sítích může být předpovídanou hranou přátelství, které je velmi pravděpodobné, ale dosud neexistuje. Predikce hrany může být využita uživateli k nalezení nových přátel [4,5].

Vzhledem k tomu, že existuje mnoho modelů používaných pro předpověď budoucí hrany v grafu, není snadné posoudit, který z modelů je lepší než ostatní. K určení nejvhodnějšího modelu můžeme použít metriky měřící přesnost predikce hrany v grafu. Využitím těchto metrik lze hodnotit výkonnost daných modelů [4,5].

Problém předpovědi hrany je dlouhotrvající výzvou moderních informačních věd. Nicméně, aktuálně chybí studiu komplexních grafů zvážení strukturálních charakteristik daného grafu. Patří mezi ně hierarchická organizace [6] a struktura společenství [7], které mohou poskytnout užitečné informace a poznatky pro predikci hrany v grafu. Některé přístupy, jako je např. náhodné procházení grafu nebo metody maximální pravděpodobnosti, našli v této problematice uplatnění.

2.1 Metody založené na podobnosti

Nejjednodušší metody predikce hrany jsou založené na podobnosti. Kde každé dvojici uzlů X a Y je přiřazeno skóre SXY, které je definováno jako podobnost (nebo také blízkost) mezi X a Y. Hrany, kterým se aktuálně nevěnuje pozornost, jsou seřazeny podle skóre a hrany spojující vrcholy, které jsou podobné těm sledovaným, by měly mít vyšší pravděpodobnost existence. Navzdory své jednoduchosti je v dané problematice studie metod založených na podobnosti vrcholů dosti složitá. Ve skutečnosti je určení vrcholu, který je podobný vrcholu zkoumanému netriviálním problémem. To hlavně z důvodu, že definice podobnosti může být jednoduchá nebo velmi složitá [8].

Podobnost vrcholů může být definována využitím základních atributů vrcholů. Dva vrcholy jsou považovány za podobné, pokud mají mnoho společných rysů [9]. Nicméně, atributy vrcholů jsou obvykle skryté, a tak je potřeba zaměřit se na jinou skupinu indexů podobnosti, např. na strukturální podobnost, která je založena pouze na struktuře grafu. Indexy strukturální podobnosti [10] lze klasifikovat různými způsoby: lokální index podobnosti oproti globálnímu indexu podobnosti, volný parametr oproti závislému parametru, závislost na vrcholu oproti závislosti na hraně a další. Obecně mohou indexy podobnosti být také klasifikovány jako strukturální ekvivalence nebo indexy pravidelné rovnocennosti.

Strukturální ekvivalence [10] udává podobnost mezi dvěma koncovými vrcholy. Oproti tomu pravidelná rovnocennost [11] předpokládá, že dva vrcholy jsou podobné, pokud se i jejich sousedé sobě podobají.

Veškeré indexy podobnosti můžeme rozdělit do tří skupin:

- Lokální indexy podobnosti
- Globální indexy podobnosti
- Pseudo-lokální indexy podobnosti

Ve srovnání s lokálními indexy podobnosti, mají globální indexy podobnosti informace o topologii grafu. I když mohou globální indexy podobnosti poskytnout mnohem přesnější předpovědi než lokální indexy podobnosti, mají globální indexy podobnosti dvě velké nevýhody:

- Výpočet globálního indexu je časově velmi náročný a je obvykle nepoužitelný pro velké sítě.
- V některých případech nejsou globální informace o topologii sítě k dispozici, např. pokud bychom chtěli implementovat metodu decentralizovaně, tedy rozdělit zkoumání sítě do více menších skupin.

Slibným kompromisem jsou pseudo-lokální indexy podobnosti, které zohledňují více informací než lokální indexy a zároveň dokážou vynechat nadbytečné informace, které přispívají k přesnosti predikce buď velmi málo, nebo vůbec.

2.2 Metody založené na maximální pravděpodobnosti

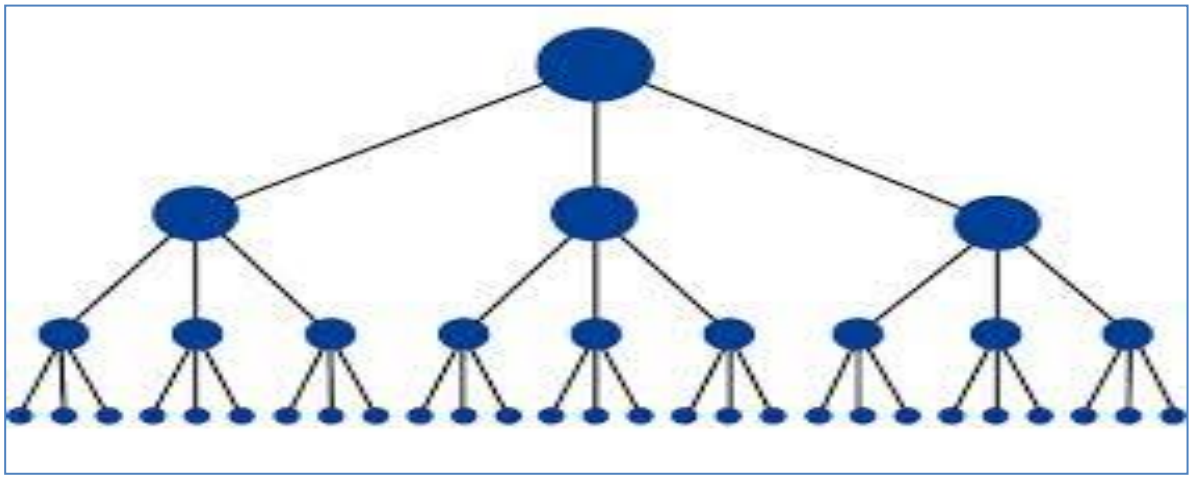
Metody založené na odhadu maximální pravděpodobnosti předpokládají organizační principy struktury grafu, která má svá podrobná pravidla a specifické vlastnosti. Pravděpodobnost jakéhokoliv nezpozorovaného spojení (spojení, o kterém doposud nebylo zjištěno, že existuje), lze vypočítat podle daných pravidel a parametrů struktury tohoto grafu. Ze známých experimentů s metodami pro predikci hrany v grafu je zřejmou nevýhodou metod maximální pravděpodobnosti to, že jsou časově velmi náročné. Dobře navržená metoda pro predikci hrany v grafu je schopna zanalyzovat graf až několika tisíců vrcholů ve srovnatelné době, jako to zvládnou jiné metody pro predikci budoucí hrany v grafu. Oproti tomu např. analýza obrovských on-line grafů (sítí), které se často skládají až z miliónů vrcholů, by se časově nevyplatila. Metody založené na maximální pravděpodobnosti nepatří mezi ty nejpřesnější, ale poskytují velmi cenné postřehy o organizaci grafu. Jedná se o postřehy, které nemohou být získány z metod založených na podobnosti.

2.2.1 Model hierarchické struktury

Mnoho reálných grafů je organizováno hierarchicky [6] - vrcholy jsou rozděleny do skupin, dále dělených do podskupin atd. (viz Obrázek 1)[1]. Dle Rednera [12], zaměřením se na hierarchickou strukturu, která je vlastní sociálním a biologickým grafům, je možné poskytnout způsob, jak najít chybějící hrany.

[1] Dental Marketing and Dentist Website Design [online]. [cit. 21.1.2014]. Dostupné z: *http://www.dental-media.co.uk/marketing/*

Ještě větším přínosem, než je predikce hrany v grafu, je, že model hierarchické struktury odkrývá skryté hierarchické uspořádání grafu. Nicméně, jak bylo uvedeno výše, velkou nevýhodou je, že tato metoda často pracuje velmi pomalu. Pro srovnání, v závislosti na procesoru moderního stolního počítače, modelem hierarchické struktury nelze analyzovat graf desítek tisíc vrcholů, zatímco algoritmy založené na lokálních indexech podobnosti mohou analyzovat grafy až s desítkami milionů vrcholů. Dalším z problémů tohoto modelu je, že může poskytnout špatné předpovědi u grafů bez jasných hierarchických struktur.



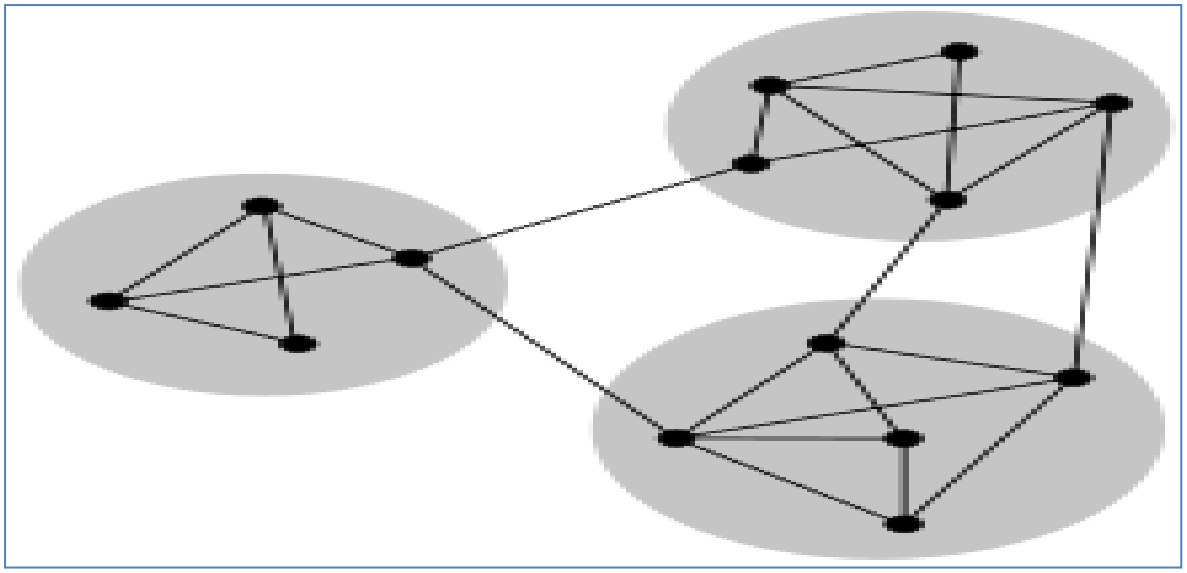
Obrázek 1 - Ukázka modelu hierarchické struktury

2.2.2 Stochastický blokový model

Stochastický blokový model [13-16] (viz. Obrázek 2)[2] je jedním z nejvíce obecných modelů grafu. Zde jsou vrcholy rozděleny do skupin a pravděpodobnost, že dva vrcholy jsou spojené, záleží pouze na skupinách, do kterých patří. Členství ve skupině hraje významnou roli při určování, jak vrcholy komunikují s ostatními. Celý proces je výpočetně velmi náročný. Tato metoda umožňuje spravovat grafy složené pouze z několika tisíců vrcholů, což v dnešní době není dostačující.

Empirické srovnání na pěti různorodých grafech (sociální interakce v klubu karate [17]; sociální graf 62 delfínů, kteří se často shlukují [18]; letecké dopravní grafy ve východní Evropě [19]; neuronový graf hlístice *Caenorhabditis elegant* [20]; metabolický graf *Escherichia coli* [21]) ukázala, že celkový výkon metody maximální pravděpodobnosti založené na stochastickém blokovém modelu [22] je lepší než výkon modelu využívajícího hierarchické struktury nebo výkon metody založené na podobnosti.

[2] Community structure – Wikipedia, the free encyclopedia [online]. [cit. 21.1.2014]. Dostupné z: *http://en.wikipedia.org/wiki/Community_structure*



Obrázek 2 - Ukázka Stochastického blokového modelu.

2.3 Pravděpodobnostní modely

Pravděpodobnostní modely se zaměřují na konkretizaci základní struktury pozorovaného grafu a předpovídají chybějící hrany pomocí naučeného modelu. Pravdivostní model grafu analyzuje graf z určitých parametrů, které mohou být nejvhodnějšími zjištěnými údaji. Jsou tři hlavní modely, respektive tzv. pravděpodobnostní relační model *(Probabilistic Relational Model)* [23], pravděpodobnostní objektově-vztahový model *(Probabilistic Entity Relationship Model)* [24] a stochastický relační model *(Stochastic Relational Model)* [25].

2.4 Použití metod pro predikci vzniku nové hrany

Problém predikce hrany přitahuje hodně pozornosti různorodých výzkumných komunit. To je přičítáno především jeho široké použitelnosti. V některých grafech, zejména biologických, metabolických a potravinových, je objevení hrany nákladné jak teoreticky, tak reálně. Přesná předpověď proto může snížit experimentální náklady a urychlit tempo výzkumu [12,26]. Dalšími možnostmi použití je např. předpověď, zda konkrétní herec bude hrát v konkrétní hře [27], predikce spolupráce v grafu spoluautorství [28], detekce vztahů mezi teroristy [26] atd.

3 Algoritmus PageRank

Algoritmus PageRank [29] byl vyvinut s cílem určení významnosti webových stránek. Podle této významnosti se následně řadí relevantní stránky ve výsledcích vyhledávačů. Tento způsob využívá např. *Google.com*. Jak autoři algoritmu uvádějí, koncept vychází z citační analýzy, kde je využíváno referencí v podobě citací. V prostředí Internetu je při určování významnosti webové stránky využíváno hypertextových odkazů, které na stránku odkazují, a významnosti webů, ze kterých tyto odkazy vedou. Algoritmus je iterativní, to znamená, že při výpočtu je použito více iterací. Z matematického pohledu je vyhodnocován graf, kde vrcholy jsou webové stránky a hrany vyjadřují, že z jedné webové stránky odkazuje hypertextový odkaz na stránku jinou. Při vyhodnocování Webu se nepoužívají tzv. interní hypertextové odkazy, tj. ty odkazy, které odkazují na stránku, kde se sami nacházejí.

3.1 Matematický zápis

Algoritmus PageRank lze popsat dvěma způsoby, přičemž každý z nich je užitečný pro něco jiného. Prvním ze způsobů popisu je maticový zápis, využívaný především pro matematické zkoumání algoritmu (např. konvergence, urychlení výpočtu). Druhý způsob je užitečný především pro snazší pochopení a implementaci, jedná se o popis s využitím zápisu výpočtu pro jeden prvek.

Page a Brin ve své práci o PageRanku [30] uvedli vzorec, který ale nebyl příliš použitelný. Jednalo se o vzorec (1), kde $PR_x(A)$ je skóre PageRanku (dále jen PageRank) vrcholu A v iteraci x, U je množina všech vrcholů odkazujících na vrchol A a N_u je počet výstupních hran vrcholu u.

$$PR_{x+1}(A) = \sum_{u \in U} \frac{PR_x(u)}{N_u} \tag{1}$$

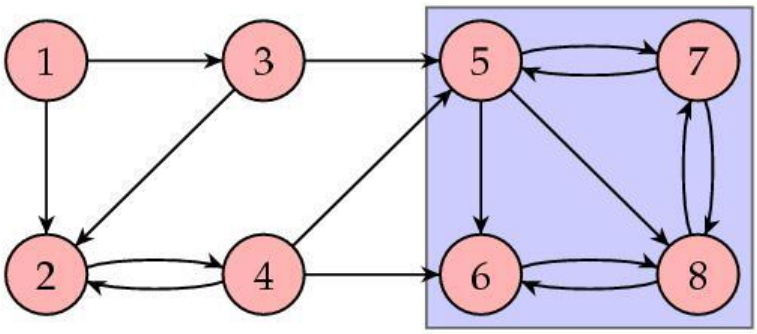
Součet PageRanku všech vrcholů grafu je roven 1, tj. 100%. Když budeme uvažovat síť Internetu, tak hodnota PageRanku udává pravděpodobnost, s jakou se webový surfař (uživatel, který prochází webové stránky umístěné na internetu), který se pohybuje po síti využitím hypertextových odkazů (hrany grafu), dostane na danou webovou stránku.

Jedním z problémů, který nastává u tohoto vzorce, je problém tzv. dangling nodes, tj. vrcholů bez jakýchkoliv výstupních hran. Problémem je, že u těchto vrcholů dochází ke ztrátě hodnoty PageRanku a poté už není součet PageRanku všech vrcholů grafu roven 1. Řešení je možné třemi způsoby [31]:

1) Vytvoření stoku – V grafu vytvoříme nový vrchol (stok), který bude mít hranu (smyčku), která ukazuje právě na něj. Poté vytvoříme všem dangling nodes výstupní hranu ukazující na tento vrchol.
2) Normalizace – Po každé iteraci normalizujeme hodnoty PageRanku všech vrcholů, tak aby jejich součet byl roven 1.
3) Rovnoměrné rozdělení – Každému dangling nodes přidáme výstupní hrany ukazující na všechny vrcholy grafu i na sebe sama.

Častěji používaným způsobem je rovnoměrné rozdělení, kde ale nemusíme přímo přidávat každému dangling nodes výstupní hrany ukazující na všechny vrcholy, ale stačí s nimi pouze počítat. První způsob není příliš vhodný, protože může dojít k situaci, že nově vytvořený vrchol (stok) může získat veškerou hodnotu PageRanku. Druhý způsob se také příliš nepoužívá a to zejména s ohledem na jeho spravedlnost. Při normalizaci je totiž každému z vrcholů přidána jiná hodnota, což může zvýhodnit některé vrcholy.

Dalším z problémů je tzv. Rank sink (klesání hodnocení), který vzniká, když vrcholy uvnitř skupiny ukazují pouze sami na sebe, ale neukazují ven ze skupiny, i když na skupinu je odkazováno z vnější (viz Obrázek 3)[3].



Obrázek 3 - Rank sink - klesání hodnocení

Pro řešení problému byl navržen model náhodného webového surfaře. Autory bylo sledováním reálných webových uživatelů zjištěno, že tito uživatelé jednou za 7 kroků při surfování na Internetu používají pro přechod na další stránku tzv. teleport, tj. napíše přímo do webového prohlížeče URL adresu webu, na který chtějí přejít. Ve zbylých 6 krocích po Internetu surfují pomocí odkazů na webových stránkách. Model náhodného surfaře byl do algoritmu PageRank vložen konstantou zvanou *damping faktor*. Užití teleportu je tedy s pravděpodobností 15%, proto je *damping faktor* většinou nastaven na 0,85, ale samozřejmě se může měnit, podle toho jestli chceme klást větší důraz na hrany grafu nebo na personalizaci (zvýhodnění vrcholu). Čím více se *damping faktor* blíží k jedné, tím větší počet iterací je nutný k dosažení zvolené přesnosti, navíc je kladen větší důraz na hrany grafu. V opačném případě je větší důraz kladen na personalizaci vrcholu.

Poslední nepřesností je, že každý hypertextový odkaz má ve výpočtu stejnou váhu, tj. pokud je na stránce 5 hypertextových odkazů, tak předpokládáme, že každý z nich bude využit s pravděpodobností 1/5. Pro zvýhodnění některého z odkazů bychom museli vzorec algoritmu doplnit o váhy jednotlivých odkazů.

3.2 Vzorec použitý v této práci

V některých případech je požadováno, aby některé z vrcholů byly v průběhu výpočtu zvýhodněny. Toho lze docílit tzv. personalizací, jedná se o zvýhodnění některých vrcholů s ohledem na jejich vlastnosti, potřeby atd.

[3] Checker PageRank [online]. [cit. 20.4.2014]. Dostupné z: *http://checkerpagerank.com/algorit.php*

Z veškerých provedených úprav vznikl vzorec (2) [31], který je použit pro tuto práci.

$$PR_{x+1}(A) = \frac{(1-d)\cdot P_A}{\sum_{p\in P} p} + d\cdot\left(\sum_{u\in U}\frac{PR_x(u)\cdot w_{u\text{toA}}}{w_{u\text{out}}} + \frac{1}{|V|}\sum_{s\in D} PR_x(s)\right) \qquad (2)$$

Ve vzorci (2) je $PR_{x+1}(A)$ je hodnota PageRanku pro jednu danou iteraci, *d* je *damping faktor*, který je pro tuto práci nastaven na hodnotu 0,85, P_A je personalizace vrcholu *A*, *P* je vektor personalizací, $w_{u\text{toA}}$ je váha hrany vedoucí z vrcholu *u* do vrcholu *A*, $w_{u\text{out}}$ je součet vah všech výstupních hran z vrcholu *u*, $|V|$ je počet všech vrcholů grafu a *D* je množina *dangling nodes*.

4 Sporty a konkrétní soutěže použité pro predikci vítězů

Z velkého množství sportů jsem zvolil ty sporty, které jsou ve světě nejvíce rozšířené. Těmito sporty jsou dle mého pohledu fotbal, hokej a tenis.

4.1 Fotbal

Zde budu zkoumat nejvyšší anglickou ligu, která nese název *Barclays Premier League*. Tuto soutěž jsem si vybral, jelikož je dle mého hlediska nejlepší na světě a dost se o ni zajímám. *Barclays Premier League* hraje 20 mužstev a každé mužstvo se utká se všemi ostatními celkem dvakrát. *Gambrinus ligu,* která je nejvyšší českou soutěží, jsem si nevybral z důvodu, že v poslední době je nemálo spojována s korupčními aférami.

Pro srovnání, jestli použitá metoda funguje i u neprofesionálních soutěží, jsem si vybral jednu z českých regionálních soutěží a to *I.A Třídu Karlovarského kraje*. Pro tuto soutěž jsem se rozhodl, jelikož v této soutěži fotbal amatérsky provozuji a chtěl bych na této soutěži danou metodu vyzkoušet. I.A Třídu Karlovarského kraje hraje 14 mužstev.

Z tohoto důvodu, že tyto dvě předešlé soutěže mají totožnou strukturu a průběh, byla vybrána ještě jedna soutěž, která se právě strukturou a průběhem liší od dvou soutěží popsaných výše. Touto soutěží je *Liga mistrů*, kterou by si chtěl zahrát snad každý malý fotbalista, protože je nejprestižnější evropskou soutěží pořádanou Evropskou fotbalovou asociací (UEFA) pro nejlepší týmy evropských lig. Do *Ligy mistrů* se kvalifikují týmy, které se umístí na nejvyšších příčkách ve svých domácích soutěžích, např. ten, kdo vyhraje českou nejvyšší soutěž *Gambrinus ligu,* se kvalifikuje do předkola této prestižní soutěže. Každá země má různý počet míst pro kvalifikaci, tento počet záleží na koeficientu dané země u mezinárodní asociace FIFA, který je dán podle toho, jak je daná země úspěšná na mezinárodních turnajích (např. Mistrovství světa) a také v evropských soutěžích, jako je např. právě *Liga mistrů*.

V této práci se nebudeme zabývat předkolem této soutěže, ale zaměříme se na hlavní část, která je rozdělena na dvě fáze. První fáze se odehrává v osmi čtyřčlenných skupinách, kde hrají každý s každým dvakrát (na hřišti domácím a soupeřově). Z těchto skupin postupují dva týmy, které získají nejvíce bodů, do druhé fáze, která se hraje vyřazovacím způsobem na dva zápasy (opět na hřišti domácím a soupeřově). Týmy, které skončí v základních skupinách na prvních místech, hrají v následující fázi s týmy ze druhých pozic. V této fázi postupuje vždy tým s lepším součtem skóre z obou zápasů (při rovnosti rozhoduje vyšší počet branek

vstřelených na hřišti soupeře, při rovnosti obou skóre se pak prodlužuje a případně rozhodují pokutové kopy). Finále se hraje na jeden zápas na předně určeném stadiónu, aby nebyl zvýhodněn ani jeden z týmů. Zápas o třetí místo se nehraje.

Struktura a průběh *Ligy mistrů*, by mohly vést k rozdílným výsledkům predikce oproti dvěma předcházejícím soutěžím, což může posloužit k pozdějšímu porovnání získaných výsledků.

4.2 Hokej

V tomto odvětví jsem si pro zkoumání vybral nejprestižnější hokejovou soutěž na světě, kterou je americká NHL (*National Hockey League*). Tuto volbu ovlivnil fakt, že NHL je soutěží, ve které mezi sebou mužstva odehrají nejvíce zápasů za sezónu na světě. NHL v současnosti hraje 30 mužstev z USA a Kanady. Dalším faktem je rozdělení soutěže na *Západní* a *Východní konferenci*. Západní konference se navíc dále dělí na *Centrální* a *Pacifickou divizi a* Východní konference je rozdělena na *Atlantickou* a *Metropolitní divizi*. Ve Východní konferenci obsahuje každá divize osm týmů, v Západní sedm týmů. Celá sezóna je navíc rozdělena do dvou částí. První z těchto částí je tzv. Základní část, v níž každé mužstvo odehraje 82 zápasů. Po skončení základní části nastává část druhá, tzv. *play-off*. Do *play-off* postupuje osm nejlepších týmů z každé konference. Poté se hrají série na čtyři vítězná utkání. První mužstvo Východní konference hraje s osmým mužstvem Východní konference, druhé hraje se sedmým atd. Stejný systém *play-off* je i v Západní konferenci. Na konci zbudou dvě mužstva (jedno z Východní a druhé ze Západní konference), která mezi sebou sehrají finálovou sérii na čtyři vítězná utkání, ze které vzejde celkový vítěz sezóny.

Ke srovnání v tomto sportovním odvětví využiji českou nejvyšší soutěž, kterou je *Extraliga*. *Extraliga* není tolik rozsáhlá jako NHL. Rozdílem je to, že v Extralize je „pouze“ 16 mužstev a není zde žádné rozdělení do konferencí a následných divizí, ani nic podobného. Extraliga má oproti NHL tři části: Základní část, rozšířené *play-off* a *play-out*. V rozšířeném *play-off* se nejprve hraje předkolo *play-off*, ve kterém se utkají sedmý s desátým a osmý s devátým umístěným mužstvem v tabulce po základní části. Hraje se na 3 vítězné zápasy, všechny ostatní série se hrají na 4 vítězné zápasy. Poté následuje klasické *play-off*. Nejprve se hrají čtvrtfinálové série, kde hraje první s jedním z týmů, který postoupí z předkola, druhý s druhým týmem, který postoupí z předkola, třetí s šestým a čtvrtý s pátým. Ze čtvrtfinálových sérií postoupí 4 mužstva, která spolu sehrají dvě semifinálové série a vítězní semifinalisté postoupí do finále. Vítěz finále se stává Mistrem extraligy. Jedenácté až čtrnácté mužstvo tabulky po základní části hrají *play-out* skupinu o udržení, ve které se každý s každým utká dvakrát, přičemž se započítávají body ze základní části. Týmy, které skončí v *play-out* skupině na posledních dvou místech, hrají baráž o udržení v extralize.

Struktura obou těchto soutěží by mohla přinést zajímavé výsledky predikce a jejich následné srovnání.

4.3 Tenis

Toto sportovní odvětví jsem si vybral zejména proto, že úspěch záleží pouze na výkonu jednotlivce, kdežto ve dvou předcházejících odvětvích záleží na kolektivním výkonu celého týmu. Pokud by se např. zranil klíčový hráč mužstva, tak by to mohlo ovlivnit celkový

výsledek. Což se v tenise stát nemůže, jelikož je každý odkázán sám na sebe. A proto by v tomto odvětví mohli být lepší výsledky predikce budoucích hran. V tenise existuje celosvětový žebříček. Veškeré zápolení mezi tenisty probíhá turnajovým systémem. Práce je zaměřena hlavně na Grandslamové turnaje (*Australian Open, French Open, Wimbledon, US Open)*, které jsou neprestižnější. K těmto čtyřem hlavním turnajům přidám tzv. Turnaj mistrů, kde hraje vždy 8 aktuálně nejlépe postavených tenistů v celosvětovém žebříčku.

Pro srovnání této oblasti byla vybrána dosti podobná soutěž, nicméně by mohla k následnému dobrému srovnání výsledků. Tato soutěž probíhá opět na Grandslamových turnajích (*Australian Open, French Open, Wimbledon, US Open)* a také na *Turnaji mistrů*, ale zde se nezaměříme na tzv. *single*, tenis jeden proti jednomu, ale na tzv. *double*, tedy tenis dvou tenistů proti dvěma. Důvodem zkoumání této soutěže je, že už nezáleží pouze na schopnostech jednotlivce, ale obou tenistech, čímž by mohly být získány dobré výsledky pro srovnání úspěšnosti predikce v tomto odvětví.

4.4 Házená

Pro výběr tohoto sportovního odvětví jsem se rozhodl hlavně proto, že zde padá větší počet gólů, což by mohlo hrát zajímavou roli, při predikci budoucí vítězů v typech grafů, kde jsou vahami hran právě vstřelené góly (viz kapitola 5).

V tomto odvětví jsem se rozhodl vybrat jednu z nejvíce prestižních soutěží na světě, kterou je německá *Bundesliga*, kde hraje i nejlepší český házenkář Filip Jícha. Tuto soutěž hraje 18 týmů, které spolu hrají způsobem každý s každým dvakrát, z čeho vyplívá, že soutěž má 34 kol. Vítězí tým, který má po 34. kole nejvíce bodů.

Pro srovnání v tomto odvětví jsem si opět vybral soutěž s jinou strukturou a průběhem než je soutěž předešlá. Tyto kritéria splňuje házenkářská *Liga mistrů*, která je obdobou fotbalové *Ligy mistrů* (viz část 4.1). Opět se nebudeme zabývat kvalifikací, ale zaměříme se na hlavní část. Ta je rozdělena na tři fáze, hlavní fáze se odehrává ve čtyřech šesti členných skupinách, kde hraje každý s každým dvakrát (v domácí a soupeřově hale). Z každé této skupiny postupují čtyři týmy do druhé fáze, která se hraje vyřazovacím způsobem na dva zápasy (opět v domácí a soupeřově hale). Hrají spolu týmy, které se v dané skupině umístily na první pozicí, s týmy z jiných skupin umístěných na 4., posledních postupových pozicích. Vždy postupuje tým s lepším součtem skóre z obou zápasů (při rovnosti rozhoduje vyšší počet branek vstřelených na hřišti soupeře, při rovnosti obou skóre se pak prodlužuje a případně rozhodují trestné hody). Poslední fází je tzv. Final Four, kde se hraje semifinále, finále a zápas o třetí místo, vše v této fázi je hráno vyřazovacím způsobem na jeden zápas.

I v tomto odvětví by rozdílná struktura a průběh dvou zkoumaných soutěží mohla vést k zajímavému srovnání úspěšnosti predikce.

4.5 Basketbal

Toto odvětví by také mohlo posloužit k lepším výsledkům predikce, při použití typů grafů, kde jsou vahami hran naházené body (viz kapitola 5), stejně jako u předchozího odvětví, tak i zde získávají hodně bodů, dokonce se počet bodů velmi často přehoupne přes 60 bodovou

hranici, v některých případech i přes 100 bodovou hranici. Ale ještě lepší výsledky predikce by mohli nastat u typu grafu, kde jsou váhy určené podle rozdílu výsledného skóre., protože v basketbalu často nastává velký rozdíl výsledného skóre, jaký nenastane v žádném jiném ze zkoumaných odvětví.

U tohoto odvětví byly opět vybrány dvě soutěže s rozdílnými strukturami a průběhy, což by mohlo vést k rozdílným výsledkům, které budou dále porovnány.

První z vybraných soutěží je basketbalová *Euroliga* mužů, která je ekvivalentem fotbalové a házenkářské *Ligy mistrů*. Opět se při zkoumání úspěšnosti predikce budeme soustředit na hlavní část, která je rozdělena do tří fází. Nejprve je Hlavní fáze, která se odehrává ve čtyřech šestičlenných skupinách, se hraje způsobem každý s každým dvakrát (v domácí a soupeřově hale). Z každé skupiny postupují 4 týmy, s nejvyšším počtem bodů do fáze zvané Top 16, tato fáze se odehrává ve dvou osmičlenných skupinách, kde se opět hraje způsobem každý s každým dvakrát (v domácí a soupeřově hale). Z každé z těchto dvou skupin postupují 4 týmy do třetí fáze zvané Play Off. Zde hraje tým, který se umístil v první skupině na prvním místě s týmem, který se ve druhé skupině umístil na 4., posledním postupovém místě. Tímto způsobem kříže jsou vytvořené dvojice pro Play Off, kde se hraje vyřazovacím způsobem na tři vítězná utkání Poslední fází, kam se probojují jen čtyři nejlepší týmy z předchozích fází je Final Four, kde se odehraje semifinále, následně finále a zápas o třetí místo. Vše se hraje na jeden vítězný zápas.

Pro srovnání v tomto odvětví byla vybrána česká nejvyšší basketbalová soutěž *Mattoni NBL*, která je rozdělena na dvě části. Nejprve se hraje hlavní část, obsahující 12 týmů, které hrají způsobem každý s každým čtyřikrát. Do druhé fáze zvané Play Off postupuje osm nejlepších týmů ze základní části. První částí Play Off je čtvrtfinále, kde se utká první tým s osmým, druhý se sedmým, třetí s šestým a čtvrtý s pátým, podle toho jak se umístili v základní části. Hraje se na 3 vítězná utkání, postupující sehrají semifinálové série hrané také na 3 vítězná utkání. Vítězové semifinále svedou souboj o titul ve finálové sérii, která je hraná také na 3 vítězné zápasy. Týmy, které v semifinálové sérii neuspěly, spolu sehrají sérii o 3. místo, hranou na 2 vítězné zápasy.

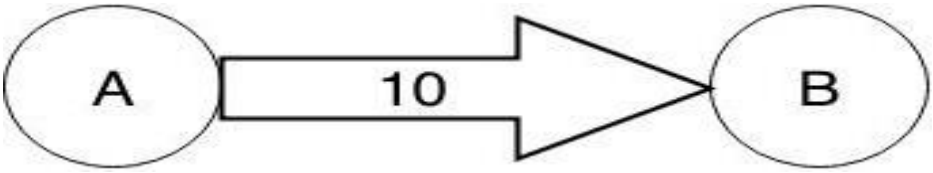
5 Typy Grafů

Pro zkoumání aplikovatelnosti algoritmu PageRank v oblasti predikce vítěze sportovního utkání byly navrženy dva typy grafů, které se od sebe liší rozdílnými vahami hran, nacházejícími se mezi vrcholy. V následující části budeme uvažovat, že vrcholy jsou mužstva (fotbal, hokej, házená, basketbal) nebo hráči (tenis).

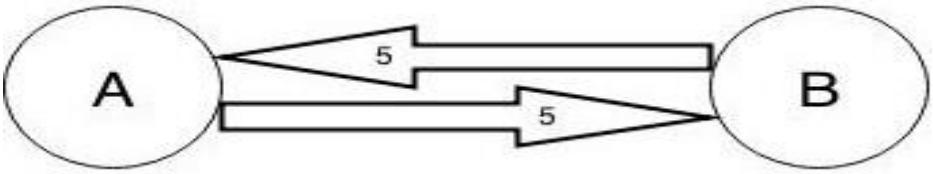
5.1 Vítězství, prohra a remíza

Tento graf je vytvořen podle toho, jestli daný tým vyhrál, prohrál nebo v zápase nastala remíza. Představme si, že zápas se odehrává mezi vrcholem A a vrcholem B. U tohoto typu grafu pokud vyhraje vrchol A, tak se vytvoří hrana vedoucí z vrcholu B do vrcholu A, kde se její váha rovná 10. Pokud by vyhrál vrchol B tak by hrana vedla opačným směrem, tedy z vrcholu A do vrcholu B se stejnou vahou (viz Obrázek 4). Pokud nastane remíza, tak se

vytvoří dvě hrany, každá s vahou 5. Jedna z těchto hran povede z vrcholu A do vrcholu B a druhá hrana povede opačným směrem (viz Obrázek 5). Jedná se o to, že pokud vyhraje vrchol A, tak ho vrchol B zvýhodňuje tím, že je mezi vrcholy vytvořena hrana, vedoucí z vrcholu, který zápas prohrál do vrcholu vítězného, s danou vahou.



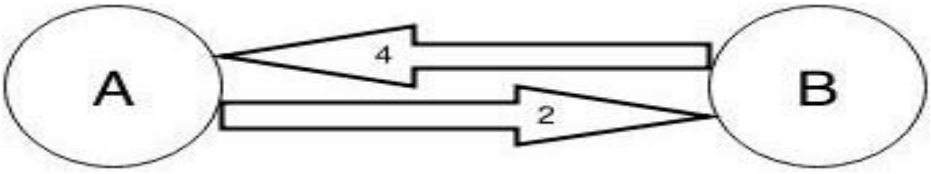
Obrázek 4 - Vítězství vrcholu B



Obrázek 5 - Remíza mezi vrcholem A a vrcholem B

5.2 Počty vstřelených gólů / získaných setů / získaných bodů

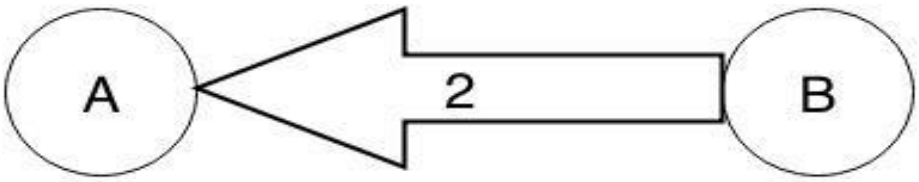
Tento graf je vytvořen podle počtu vstřelených gólů (fotbal, hokej, házená), získaných setů (tenis) nebo získaných bodů (basketbal). Pokud se odehrává zápas mezi vrcholem A a vrcholem B, který dopadne výsledkem např. 4:2, tak se graf tvoří způsobem, že je vytvořena hrana vedoucí z vrcholu A do vrcholu B s vahou 2. Zároveň je vytvořena hrana vedoucí opačným směrem s vahou 4. Jde o to, že do vrcholu vede vstupní hrana s vahou rovnou počtu vstřelených gólů daným vrcholem. Takže pokud vrchol A vstřelil 4 góly, vytvoří se hrana, vedoucí z vrcholu B do vrcholu A, s vahou 4 (viz Obrázek 6).



Obrázek 6 - Vítězství vrcholu A nad vrcholem B výsledkem 4:2

5.3 Rozdíl skoré

Tento typ grafu je vytvořen podobným způsobem jako předcházející typ grafu. Představme si situaci, kdy se daný zápas odehrává mezi vrcholem A a vrcholem B, výsledkem tohoto zápasu je např. 4:2. Graf se tvoří způsobem, že je vytvořena pouze jedna hrana vedoucí z vrcholu B do vrcholu A s vahou 2, což je rozdíl skoré výsledku daného sportovního utkání. Je tedy zvýhodněn vrchol A, protože vyhrál rozdílem 2 gólů (viz Obrázek 7).



Obrázek 7 - Vítězství vrcholu A nad vrcholem B výsledkem 4:2

5.4 Úprava personalizačního vektoru PageRanku

Personalizační vektor PageRanku byl upraven tím způsobem, že jako personalizace je jednotlivým vrcholům grafu přidělena hodnota podle toho jak se umístili v soutěži v předchozím ročníku. Vítězi minulého ročníku je přidělena nejvyšší hodnota, která se snižuje společně s nižším a nižším umístěním v soutěži v minulém ročníku až po nejnižší hodnotu, pro tým, který skončil na posledním místě, respektive nově postoupil z nižší soutěže (viz Tabulka 1).

Tabulka 1 - Ukázka přidělení hodnoty personalizace vrcholům

Tým	Umístění z předchozího ročníku	Hodnota personalizace
Manchaster Utd	1.	5
Manchaster City	2.	4
Chelsea	3.	3
Arsenal	4.	2
Tottenham	5.	1

6 Aplikace pro vytvoření (navržených) grafů a jejich vyhodnocení PageRankem

Aplikace, která slouží k vytvoření grafu, jeho následnému vyhodnocení navrženými úpravami PageRanku a výpisu výsledků do výstupního souboru.

6.1 Predikce budoucí hrany podle aktuálního postavení v tabulce

Aplikace obsahuje funkci, která predikuje vítěze sportovního utkání podle aktuálního postavení v tabulce, respektive podle počtu doposud získaných bodů. Pokud uživatel chce předpovídat vítěze podle aktuálního postavení týmů v tabulce, tak program v průběhu svého běhu tvoří tabulku, respektive přičítá týmům body podle toho, jestli vyhráli, remizovali nebo prohráli. Pokud tým vyhraje tak mu přičte 3 body, při remíze se oběma hrajícím týmům v daném zápase přičte 1 bod. Týmu, který prohraje, se nepřičítá žádný bod. Následná predikce budoucího vítěze probíhá způsobem, že aplikace zjistí kdo z týmů, které spolu mají hrát, má více bodů. Tým, který má více bodů, je automaticky předpovězen jako vítěz. Pokud mají v tabulce stejně bodů, tak je predikována remíza.

6.2 Struktura vstupního souboru

Vstupní soubor pro danou aplikaci musí být typu CSV (oddělený středníkem). Tento soubor musí mít pevně danou strukturu. Nejprve musí být na začátku vypsány všechny vrcholy (konkrétně jejich jedinečné názvy), které bude graf obsahovat. Tyto vrcholy musí být vypsané pod sebe. Pokud chceme při výpočtu predikce využít jakékoliv personalizace, tak je nutné její hodnotu uvést ve druhé buňce stejného řádku, jako vrchol, pro který je daná personalizace určena. Po vrcholech následuje řádka, kde se v první buňce nachází klíčové slovo *Start*, které slouží k zastavení načítání nových vrcholů (viz Obrázek 8). Po této řádce následuje prázdný řádek a poté začíná výpis zápasů pro první kolo soutěže, ze kterého je počítána predikce. Výpis zápasů prvního kolo je do té doby, než nastane řádek, kde je v první buňce klíčové slovo *Kolo*, které značí konec daného kola. Poté následuje prázdný řádek a dále výpis zápasů pro následující kolo (viz Obrázek 9). Tímto způsobem je strukturován celý zbytek vstupního souboru a aplikace ho čte do té doby, než se v první buňce řádky nenachází klíčové slovo *Konec*. Toto slovo značí konec vstupního souboru.

- Struktura řádky jednoho zápasu daného kola
    - *Číslo kola;Domácí tým;Hostující tým;Výsledek v řádné hrací době;Výsledek po prodloužení* (viz Obrázek 9)

Tottenham	15		
West Brom	4		
West Ham	8		
Start			
1	Liverpool	Stoke	1:0
1	Arsenal	Aston Villa	1:3
1	West Brom	Southampton	0:1

Obrázek 8 - 1. Ukázka vstupního souboru

1	Chelsea	Hull	2:0
1	Man. City	Newcastle	4:0
Kolo			
2	Fulham	Arsenal	1:3
2	Stoke	Crystal P.	2:1

Obrázek 9 - 2. Ukázka vstupního souboru

6.3 Aplikace

Aplikace se nejprve zeptá, uživatele jakým způsobem chce predikovat vítěze jednotlivých sportovních utkání (podle hodnoty PageRanku nebo podle aktuálního postavení v tabulce). V případě výběru predikce vítěze podle hodnoty PageRanku, má uživatel možnost volit jaký typ grafu se má vytvořit (viz část 4.1 a 4.2). Uživatel je také dotázán, zda chce využít personalizace jednotlivých vrcholů nebo ne.

Poté aplikace začne číst ze vstupního souboru vybraného uživatelem. Nejprve načte všechny vrcholy, které se v grafu budou nacházet. Dále každému z vrcholů nastaví počáteční hodnotu, která je rovna *1.0 / počet všech vrcholů grafu*. V případě využití personalizace přidělí hodnotu P_a dělenou sumou všech P, aby součet PageRanku všech vrcholů byl roven 1. Následuje postupné načítání jednotlivých zápasů prvního kola. Na začátku každého kola přepočítá hodnotu PageRanku pro každý vrchol. Po načtení prvního kola, aplikace ze zápasů buď vytvoří hrany mezi zúčastněnými vrcholy, nebo daným vrcholům přičte body do tabulky, to záleží na tom, podle čeho bude predikovat vítěze sportovních utkání. Dále aplikace predikuje vítěze pro daná utkání v následujícím kole, opět podle vybraného typu predikce. V prvním kole je, v případě predikce využitím právě PageRanku, ve všech zápasech predikována remíza, protože na začátku mají všechny vrcholy stejnou hodnotu PageRanku, pokud není využita personalizace vrcholů. V případě predikce využitím aktuálního postavení v tabulce mají všechny vrcholy na počátku 0 bodů, proto je i v tomto případě v prvním kole predikována remíza pro všechny zápasy.

Nakonec aplikace zapíše zápasy pro dané kolo společně s predikovanými vítězi do výstupního souboru. Tímto způsobem pracuje aplikace cyklicky až do té doby, dokud není predikován vítěz posledního zápasu posledního kola, nacházejícího se ve vstupním souboru.

6.4 Struktura výstupního souboru

Výstupní soubor vytvořený danou aplikací je typu CSV (oddělený středníkem). Tento soubor má pevnou strukturu, kterou tvoří aplikace. Na prvním řádku souboru je hlavička pro následující data, kterými jsou zápasy pro jednotlivá kola.

- Hlavička výstupního souboru
  - *Kolo;Domácí;Hosté;Výsledek v řádné hrací době;Výsledek po prodloužení;Predikce vítěze*

Data, která následují po hlavičce, mají stejnou strukturu jako hlavička. Tyto data jsou jednotlivé zápasy s reálnými výsledky a predikovanými vítězi aplikací. Jednotlivá kola jsou oddělena řádkem s klíčovým slovem *Kolo*, které určuje konec kola, jehož výčet byl právě před tímto slovem. Následuje prázdný řádek, po kterém je výpis zápasů následujícího kola. Tímto způsobem je strukturován celý výstupní soubor (viz Obrázek 10).

Kolo	Domácí	Hosté	Výsledek v řádné hrací době	Výsledek po prodloužení	Predikce
1	Liverpool	Stoke	1:0		Remiza
1	Arsenal	Aston Villa	1:3		Remiza
1	West Brom	Southampton	0:1		Remiza
1	Sunderland	Fulham	0:1		Remiza
1	Norwich	Everton	2:2		Remiza
1	West Ham	Cardiff	2:2		Remiza
1	Swansea	Man. United	1:4		Remiza
1	Crystal P.	Tottenham	0:1		Remiza
1	Chelsea	Hull	2:0		Remiza
1	Man. City	Newcastle	4:2		Remiza
Kolo					
2	Fulham	Arsenal	1:3		Fulham
2	Stoke	Crystal P.	2:1		Remiza
2	Everton	West Brom	0:0		Everton

Obrázek 10 - Ukázka struktury výstupního souboru

7 Aplikace pro vykreslení grafu úspěšnosti predikce

Tato aplikace umožňuje, vykreslení grafu úspěšnosti predikce za jednotlivá kola soutěže a možnost uložení grafu ve formátu JPEG. Také umožňuje překreslení, kdy je predikce považována za úspěšnou i v případě neprohry predikovaného vítěze (viz část 7.2).

7.1 Aplikace

Aplikace nejprve načte data ze vstupního souboru (viz část 6.2) a poté pro každý zápas zjišťuje, zda byla predikce úspěšná nebo ne. Predikce je úspěšná pokud v daném zápase vyhraje ten z vrcholů grafu, který je predikován v buňce pro predikci. V případě remízy musí být v buňce predikce *„Remíza“*. Pokud je tomu jinak tak aplikace vyhodnotí predikci vítěze jako chybnou.

Aplikace dále sčítá úspěšné predikce za dané kolo a zároveň si uchovává informaci o počtu zápasů v daném kole, protože v některých soutěžích se tato hodnota může s různými koly měnit. Po skončení aktuálně zkoumaného kola, aplikace vydělí počet úspěšných predikcí počtem zápasů v daném kole a výsledek vynásobí stem, čímž získá procentuální úspěšnost predikce vítězů pro dané kolo. Tímto způsobem zjistí procentuální predikce vítězů pro každé kolo nacházející se ve vstupním souboru a dané procentuální úspěšnosti vykreslí do grafu (viz Obrázek 11).

Obrázek 11 - Ukázka grafu, znázorňující úspěšnost predikce pro daná kola soutěže. Graf také udává informaci o průměrné úspěšnosti predikce za celou soutěž.

7.2 Úspěšnost predikce v případě neprohry

Aplikace navíc uchovává informace o úspěšnosti predikce v případě, že predikce vítěze je úspěšná i v případě neprohry vrcholu, který je predikován jako vítěz. Představme si, že sportovní klání probíhá mezi vrcholem A a vrcholem B, pokud je pro dané sportovní klání jako vítěz predikován např. vrchol A a výsledkem daného klání je vítězství vrcholu A nebo remíza, tak je predikce neprohry úspěšná. Predikce vítěze není úspěšná pouze v případě, kdy dané sportovní klání skončí vítězstvím vrcholu B.

Aplikace pro vykreslení grafu, umožňuje pomocí zaškrtávacího okénka překreslit daný graf úspěšnosti predikce na graf, kde je jako úspěšná predikce brána neprohra predikovaného vítěze (viz výše).

8 Diskuze výsledků

V této kapitole budou diskutovány získané výsledky a použití PageRanku v této oblasti. Veškeré procentuální výsledky v této kapitole jsou zaokrouhleny na 2 desetinná místa.

8.1 Diskuze výsledků a použití PageRanku ve fotbalových soutěžích

Pro zkoumání a porovnání výsledků a využití algoritmu PageRank v tomto sportovním odvětví, byly vybrány tři soutěže *Barclays* P*remier League, I.A třída Karlovarského kraje* a *Liga mistrů* (viz část 3.1).

8.1.1 Výsledky úspěšnosti predikce v případě kladné predikce při úvaze třístavového modelu (vítězství, prohra a remíza)

Výsledky při využití tohoto modelu se pohybují kolem průměrných 50%.

Výsledky *Barclays Premier League*

Výsledky predikce pro tuto soutěž jsou průměrné. Když se zaměříme na průměrné procento úspěšnosti predikce pro celou soutěž, tak se ve všech typech predikce pohybuje kolem 50% úspěšnosti.

V případě predikce využitím PageRanku, kde jsou vahami hran vstřelené góly daných vrcholů (viz část 4.2) a není zde využito personalizace, se průměrné procento úspěšnosti predikce pro celou soutěž dostává na 49,21%. Hlavním důvodem lehce podprůměrné úspěšnosti predikce by mohlo být to, že použití vstřelených gólů jako vah hran grafu, není dost dobrým řešením. Důvodem je, že v počtu střílených gólů hraje velkou roli mnoho faktorů např. styl hry daného vrcholu (útočný oproti obrannému). Vrchol může být úspěšný, i pokud vyznává obranný styl, tedy brání, aby neinkasoval žádný gól, a z ojedinělé útočně akce sám gól vstřelí. Důkazem toho tvrzení je v této zkoumané soutěži tým *Chelsea*, který praktikuje právě styl hry popsaný výše a s tímto je úspěšný.

Při použití jakéhokoliv z ostatních druhů predikce, ať už s využitím personalizace či bez personalizace vrcholů, se průměrné procento úspěšnosti predikce pro celou zkoumanou soutěž přehouplo přes 50%. Rozdíly nebyli velké, mezi nejnižší a nevyšší hodnotou byl rozdíl 2.8 procentních bodů. Druhá nejnižší hodnota byla zjištěna v případě využití predikce, kde jsou opět vahami hran vstřelené góly, nicméně zde bylo využito personalizace vrcholů (viz část 4.3), což je hlavním důvodem proč se průměrná úspěšnost predikce pro celou soutěž zvýšila o 3,16 procentních bodů na 52,37%. Není to velké zlepšení, ale je vidět, že použití personalizace vrcholů hraje malou, ale viditelnou roli, v průměrné úspěšnosti predikce.

Nejvyšší průměrná úspěšnost byla vypočítána v případě využití grafu, kde jsou váhy hran určeny podle vítězství, prohry a remízy (viz kapitola 4.1) a bylo také využito personalizace vrcholů. Zde dosáhla průměrná úspěšnost predikce pro celou soutěž 54,47%, z čehož lze

soudit, že při využití vah hran podle vítězství, prohry a remizy, je průměrná úspěšnost predikce lepší, než když jsou váhy hran rovny vstřeleným gólům. Nicméně jsem čekal, že rozdíl bude markantnější. Navíc je opět možné vypozorovat, že využití personalizace vrcholů hraje i v tomto typu grafu kladnou roli. Důkazem je nižší hodnota průměrné predikce ve stejném typu grafu, ale bez použití personalizace vrcholů (přibližně o 1,15 procentních bodů).

V případě typu grafu, kde jsou vahami hran rozdíly výsledných skóre daných sportovních utkání (viz část 5.3), se průměrná úspěšnost predikce moc neměnila. Při predikci s využitím tohoto typu grafu bez využití personalizace vrcholů, se hodnota průměrné predikce pro celou soutěž rovnala 51,58%. Využití personalizace vrcholů mělo opět kladný účinek na tuto hodnotu, jelikož se v tomto případě hodnota průměrné úspěšnosti predikce zvýšila o 3,16 procentuálních bodů na 54,47%. Tato průměrná úspěšnost predikce se přesně rovná průměrné úspěšnosti predikce s využitím personalizace vrcholů u typu grafu, kde jsou váhy hran určeny podle vítězství, prohry a remízy. Toto by mohlo způsobit to, že u obou těchto typů grafů se hrany, mezi danými vrcholy, oběma směry vyskytnou pouze v případě, že jednou zvítězí vrchol A a ve druhém případě vrchol B. U prvního typu grafu se hrany oběma směry vyskytují ještě v případě remízy (viz část 4.1), ale jak se zdá tak to u této soutěže nehrálo roli.

Překvapením byla průměrná úspěšnost predikce pro celou soutěž s využitím predikce pouze pomocí aktuálního postavení v tabulce (viz kapitola 5.1). V tomto případě dosáhla průměrná predikce hodnoty 53,68%, což je druhý nejvyšší výsledek průměrné predikce pro tuto soutěž.

Výsledky *I.A třídy Karlovarského kraje*

Výsledky predikce v této soutěži jsou lehce pod průměrnou hranicí 50%. Při zaměření se na hodnotu průměrné procentuální úspěšnosti pro celou soutěž, se tato hodnota ve všech typech predikce pohybuje mezi 41% až 50%.

U této soutěže byly nejhorší výsledky získány v případě predikce použitím PageRanku s typem grafu, kde jsou hrany ohodnoceny podle vítězství, prohry a remízy bez použití personalizace vrcholů. Zde byla hodnota průměrné procentuální úspěšnosti rovna 41,67%. V případě stejného grafu a využití personalizace vrcholů, se hodnota průměrné úspěšnosti predikce zvýšila na 44,23%. Příliš velká změna to není, ale i u této soutěže hraje personalizace vrcholu kladnou roli v predikci.

V případě využití grafu, kde jsou vahami hran vstřelené góly, byla hodnota průměrné úspěšnosti predikce podobná té z předchozího typu grafu. Důsledkem menších rozdílů u obou typů grafů by patrně mohla být vyrovnanost soutěže, vyjádřená tím, že i absolutní outsider[4] může porazit toho největšího favorita a spíše záleží na schopnosti vrcholů střílet góly, respektive na jednotlivých hráčích v jejich řadách. Protože v těchto nižších regionálních soutěžích nezáleží až tak moc na celém týmu, jako tomu je v nejlepších soutěžích na světě, ale stačí jeden vydařený zápas jednoho z hráčů a vše se může změnit. Hodnota průměrné úspěšnosti predikce, kde není využito personalizace vrcholů, se rovná 43,59%. Pokud využijeme personalizaci, tak to má opět kladný důsledek na úspěšnost výsledků predikce, což

[4] Outsider – tým, který má téměř nulovou šanci vyhrát

je vidět na hodnotě průměrné úspěšnosti predikce pro celou soutěž, která se zvýšila na 44,87%.

Pro typ grafu, kde jsou váhy hran určeny podle rozdílu výsledného skóre, mělo využití personalizace opět kladný účinek na hodnotu úspěšnosti predikce. Důkazem je hodnota průměrné úspěšnosti predikce, kdy nebylo využito personalizace, která se rovná 41,03%, oproti hodnotě průměrné úspěšnosti predikce při využití personalizace vrcholů se tato hodnota zvýšila o 4,49 procentuálních bodů na 45,51%.

Nejvyšší hodnota průměrné úspěšnosti predikce pro zkoumanou soutěž byla získána trochu překvapivě využitím predikce pouze podle aktuálního postavení v tabulce, kde hodnota dosáhla 49,36%. To, že právě tento typ predikce, je v této soutěži nejvíce úspěšný, i když jeho průměrná hodnota je stále podprůměrná, může mít opět za následek vyrovnanost této soutěže. Proto jsou vrcholy, které se nacházejí na předních pozicích tabulky (mají nejvíce bodů), zvýhodněny a je tu předpoklad, že porazí ty vrcholy nacházející se na konci dané tabulky (mají nejméně bodů).

Výsledky pro tuto soutěž byly lehce podprůměrné a nedá se tedy říci, že by bylo možné daný typ predikce budoucí hrany v této soutěži využít.

Výsledky *Ligy mistrů*

Výsledky predikce pro tuto soutěž se většinou pohybovali pod 50% procentní hranicí. Pouze v jednom typu predikce byla tato hodnotu překonána.

Nejhorších výsledky predikce pro tuto soutěž byly zjištěny u typu grafu, kde jsou váhy hran určeny podle rozdílu výsledného skóre. Tato situace nastala při predikci bez využití personalizace vrcholů. Hodnota průměrné úspěšnosti predikce byla rovna 37,98%, což je hluboko pod průměrem. Ale zajímavý výsledek přinesla úspěšnost predikce pro totožný graf s využitím personalizace vrcholů, protože v tomto případě se hodnota průměrné úspěšnosti predikce zvýšila o 10,58% na 48,56%, což je doposud největší zpozorovaný rozdíl.

Lehce podprůměrných hodnot dosahovala hodnota průměrné úspěšnosti predikce u typu grafu, kde jsou váhy hran určeny podle vítězství, prohry a remízy. Zde se opět potvrdilo to, že využití personalizace vrcholů má kladný dopad na úspěšnost predikce, jelikož v případě, kdy personalizace vrcholů využito nebylo, byla hodnota průměrné úspěšnosti predikce rovna 41,83% a při využití personalizace vrcholů se tato hodnota zvýšila na 46.63%.

Nejlepší úspěšnost predikce u této soutěže byla zjištěna u typu grafu, kde jsou vahami hran počty vstřelených gólů jednotlivých týmů. S využitím personalizace vrcholů u tohoto typu grafu byla průměrná úspěšnost predikce rovna 50,98%, v případě nevyužití personalizace vrcholů se průměrná úspěšnost predikce snížila na 47.12%. Opět je vidět kladný dopad využití personalizace vrcholů na úspěšnosti predikce.

Při využití predikce podle aktuálního postavení v tabulce byly zjištěny celkem dobré výsledky s ohledem na předchozí typy predikce. U tohoto typu predikce se průměrná hodnota úspěšnosti predikce rovnala 48,56%. Pozoruhodné je, že tato hodnota se přesně rovná hodnotě, zjištěné při predikci pomocí PageRanku s využitím personalizace vrcholů, u typu grafu, kde jsou váhy hran určeny podle rozdílu výsledného skóre.

Výsledky v této soutěži se příliš nelišili, často byly lehce podprůměrné, v jednom případě byla odchylka větší, což může být způsobeno tím, že ke druhé fázi této soutěže se přistupuje s rozdílnými taktickými úkoly a dané týmy se soustředí především na to, aby neinkasovali gól

Porovnání výsledků predikce všech tří zkoumaných soutěží tohoto sportovního odvětví

Když porovnáme výsledky všech tří zkoumaných soutěží tohoto odvětví, tak je vidět, že nejlepší výsledky byly v případě *Barclays Premier League*, ale v žádném případě se nejednalo o velký rozdíl, právě naopak. V *Barclays Premier League* dosahovala výsledná hodnota průměrné úspěšnosti predikce přibližně 50%. *V I.A třídě Karlovarského kraje*, stejně jako v *Lize mistrů* byly výsledky lehce podprůměrné, pouze v jednom z typů predikce se hodnota průměrné úspěšnosti predikce přiblížila na dosah průměrné 50% hodnoty.

Důvodem takovýchto výsledků by mohla být ta skutečnost, že *I.A třída Karlovarského kraje* a *Liga mistrů*, zejména v první fázi soutěže, nejsou ani zdaleka tolik vyrovnané jako *Barclays Premier League*. Zároveň se v nižších českých soutěžích objevuje problém ovlivňování zápasů, což určitě mohlo hrát významnou roli v konečných výsledcích.

Ze získaných výsledků je také možné vypozorovat, že v případě *Barclays Premier League* a *Lize mistrů* byly vykazovány lepší výsledky predikce u typu grafu, kde jsou hrany ohodnoceny podle vítězství, prohry a remízy. Naopak v případě zkoumání *I.A třídy Karlovarského kraje* byly lepší výsledky u druhého typu grafu, kde jsou vahami hran vstřelené góly. Toto může mít za dopad to, že v nižších regionálních soutěžích padá mnohem více gólů, protože týmy se soustředí především na útočnou fázi. Nezřídka kdy dojde i ke vstřelení dvouciferného počtu gólů. Oproti *Barclays Premier League* a zejména druhé fázi *Ligy mistrů,* kde tolik gólů nepadá.

Z dostupných získaných výsledků docházím k závěru, že tento typ predikce vítězů fotbalových utkání není v této oblasti příliš použitelný.

8.1.2 Diskuze výsledků úspěšnosti predikce v případě kladné predikce při úvaze stavu neprohry

U tohoto sportovního odvětví je rozdíl změny, při použití daného modelu v průměru o 20 až 25 procentuálních bodů (viz Obrázek 13 a Obrázek 14). Což je následkem toho, predikce remízy je minimální, ale určitá část zápasů končí právě remízou.

Obrázek 13 - Ukázka grafu úspěšnosti predikce v třístavovém modelu (vítězství, prohra a remíza)

Obrázek 14 - Ukázka grafu úspěšnosti predikce v modelu, kdy je uvažován stav neprohry

8.2 Diskuze výsledků a použití PageRanku v hokejových soutěžích

V tomto odvětví byly očekávány lepší výsledky než v odvětví fotbalu a z toho důvodu, že se soutěže v tomto odvětví hrají na více kol než v ostatních odvětvích a vrcholy se tedy mnohem více ovlivňují.

8.2.1 Výsledky úspěšnosti predikce v případě kladné predikce při úvaze třístavového modelu (vítězství, prohra a remíza)

Výsledky *NHL*

V této soutěži byly předpokládány dobré výsledky, vzhledem k počtu kol a tudíž větší interakce mezi vrcholy. Ale opak byl pravdou, hodnoty průměrné úspěšnosti se pohybovali kolem 50%. Výsledky zde byly velmi vyrovnané ve všech typech predikce.

Nejnižší průměrná hodnota úspěšnosti predikce pro celou soutěž byla zjištěna při použití typu grafu, kde jsou vahami hran počty vstřelených gólů. Nejnižší hodnota byla zjištěna, když nebylo využito personalizace vrcholů, její výše byla 41,41%. Při zohlednění personalizace vrcholů se hodnota průměrné úspěšnosti zvýšila na hodnotu 44,31%. Není to velké zlepšení,

ale opět můžeme konstatovat, že i v tomto případě pomáhá personalizace vrcholů k lepší úspěšnosti predikce.

U druhého typu grafu, kde jsou hodnoty hran určeny podle vítězství, prohry a remízy, byli hodnoty průměrné úspěšnosti predikce o trochu vyšší. V případě, kdy nebylo využito personalizace vrcholů, byla její hodnota 42,72%. Pokud u tohoto typu grafu navíc využijeme personalizaci vrcholů, pak se hodnota průměrné úspěšnosti predikce pro celou soutěž zvýší na 43,51%, což je přibližně o 0,8 procentuálních bodů nižší, oproti predikci s využitím personalizace vrcholů v předchozím typu grafu.

Změna bude zřejmě způsobena tím, že se v této zkoumané soutěži střílí hodně gólů, což u prvního typu grafu, může způsobovat větší vyrovnanost vrcholů. Samozřejmě, ve výpočtu nehraje roli pouze váha hrany, je tam plno dalších atributů, které je nutné zohlednit. Ale mou domněnkou je, že rozdíl v této oblasti je způsoben právě větší vyrovnaností vrcholů, díky velkému počtu střílených gólů.

Pro třetí typ grafu byly zjištěny obdobné výsledků jakou u dvou předcházejících typů. Personalizaci vrcholů měla opět kladný dopad na průměrnou úspěšnost predikce, s jejím využitím byl úspěšnost predikce rovna 44,55%, v případě nevyužití personalizace vrcholů klesla úspěšnost predikce na 43,53%.

Posledním typem predikce, který byl v této soutěži testován, je predikce podle aktuálního postavení v tabulce. Výsledek úspěšnosti predikce využitím této metody se nijak zvláště nelišil od výsledků zjištěných ostatními typy. Hodnota průměrné úspěšnosti predikce pro celou soutěž byla 42,52%. Hodnota je tedy opět pod průměrem. Zde to může být způsobeno tím, že zkoumaná soutěž je rozdělena do dvou větších skupin, které jsou ještě rozděleny do další dvou skupin (viz část 3.2). Proto, pokud spolu hrají týmy z různých skupin, mohou mít podobný nebo i rozdílný počet bodů, ale záleží na tom, jak spolu např. hráli předchozí vzájemný zápas, protože týmy z různých skupin se příliš nepotkávají, čímž není tak jednoduché predikovat vítěze.

Výsledky Extraligy

V této soutěži se průměrné hodnoty úspěšnosti predikce pohybovaly kolem 50%. Byly předpokládány lepší výsledky s ohledem na vyšší počet kol a tedy větší ovlivňování se vrcholů mezi sebou.

Když se opět pro porovnání zaměříme na průměrnou úspěšnost predikce pro celou soutěž, tak průměrná úspěšnost predikce zjištěná při použití typu grafu, kde hodnotami vah hran jsou počty vstřelených gólů, a není využíváno personalizace vrcholů. V tomto typu predikce bylo dosaženo průměrné úspěšnosti 45,65%. Využitím personalizace vrcholů se tato úspěšnost zvýšila na 50,97%. Je vidět, že i v tomto případě hraje personalizace vrcholu, kladnou roli při úspěšnosti predikce.

U typu grafu, kde jsou hodnoty vah hran určeny podle vítězství, prohry a remízy se hodnota průměrné úspěšnosti predikce mírně zlepšila. Personalizace vrcholů hraje i u tohoto typu grafu kladnou roli při zlepšení průměrné úspěšnosti predikce. Při predikci bez personalizace vrcholů je hodnota průměrné úspěšnosti predikce 47,81%, při využití personalizace vrcholů se

tato hodnota zvýší na 51,28%. Je vidět, že zvýhodnění vrcholu podle postavení v tabulce z minulého ročníku zlepšuje výsledky.

Při zkoumání posledního třetího typu grafu, kde jsou váhy hran určeny rozdílem výsledného skóre. V tomto případě se opět potvrzuje kladný dopad využití personalizace vrcholů na úspěšnost predikce. Bez využití personalizace vrcholů je průměrná úspěšnost predikce 44,01%, což je nejnižší hodnota úspěšnosti predikce, která byla v této soutěži zjištěna. Při využití personalizace vrcholů se hodnota průměrné úspěšnosti predikce zvýšila na 46,15%.

Celkem nízká průměrná úspěšnost predikce, konkrétně 46,93%, v této soutěži je při použití predikce podle aktuálního postavení v tabulce. To je důsledkem toho, hokejové tabulky jsou většinou dost vyrovnané a nejsou tam tak velké bodové rozdíly. Pokud tedy dělí tým, který je např. na 1. místě v tabulce a tým např. na 7. místě tabulky pouze malý počet bodů, tak není lehké predikovat vítěze.

V této soutěži jsou nejlepší výsledky průměrné predikce získány při predikci využitím typu grafu, kde jsou váhy hran určeny podle vítězství, prohry a remízy společně s personalizací vrcholů. To může být způsobeno tím, že v hokeji se střílí hodně gólů na obou stranách. Proto jsou v typu grafu, kde jsou hodnoty vah hran určeny podle počtu vstřelených gólů, dosti zvýhodňovány oba vrcholy, i když jeden z nich prohraje.

Porovnání výsledků predikce obou zkoumaných soutěží tohoto sportovního odvětví

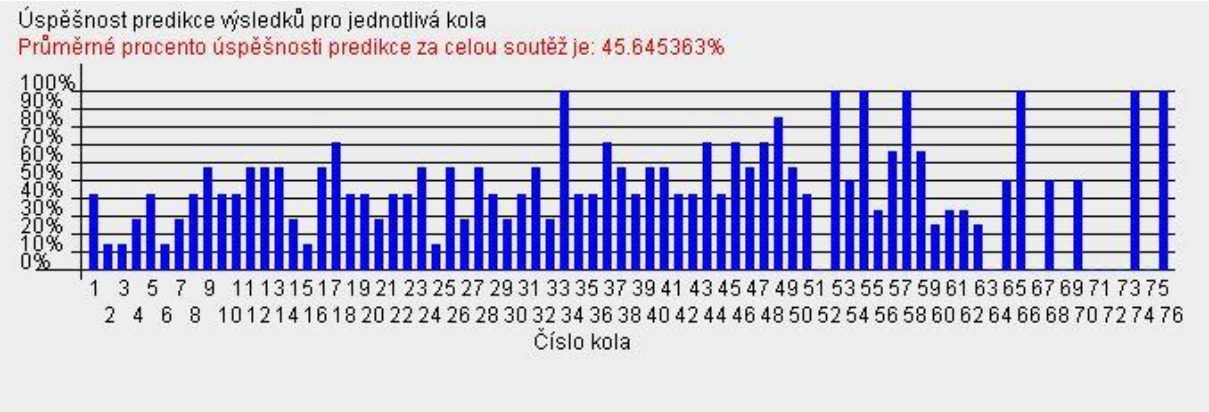
V tomto odvětví byly porovnávány dvě soutěže *NHL* a česká nejvyšší hokejová soutěž *Extraliga*. Bylo by možné předpokládat, že když se v *NHL* hraje více zápasů než v *České hokejové Extralize*, tak by i výsledky mohli v této soutěži být lepší vzhledem k většímu ovlivnění vrcholů.

Nicméně není tomu tak, hodnoty jsou téměř totožné. Malé rozdíly by se našly, ale jsou opravdu zanedbatelné. Výsledky v obou soutěžích jsou téměř shodné, což ale nebylo předpokládáno, protože byly vybrány právě tyto dvě soutěže, jelikož jsou rozdílné svou strukturou (viz kapitola 3.2), ale ani tento fakt nenapomohl k rozdílným výsledkům.

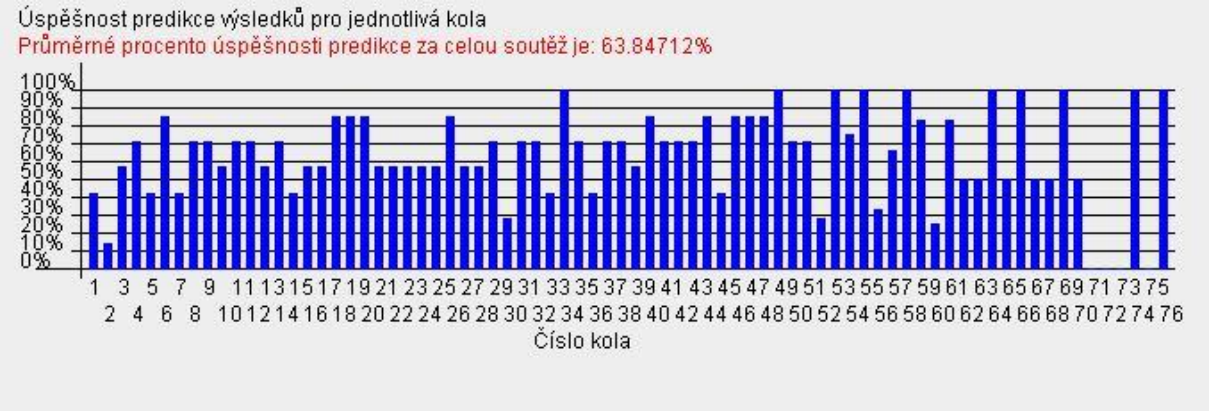
U fotbalových soutěží se nejlépe osvědčila predikce dle algoritmu PageRank pro graf, kde jsou váhy hran určeny podle vítězství, prohry a remízy. U hokejových soutěží je to podobné ale výsledky jsou ještě více vyrovnané, a to zřejmě díky většímu počtu střílených gólů

8.2.2 Diskuze výsledků úspěšnosti predikce v případě kladné predikce při úvaze stavu neprohry

V případě tohoto sportovního odvětví je změna oproti fotbalovým soutěžím menší, v průměru se jedná o 18 procentuálních bodů (viz Obrázek 15 a Obrázek 16), což by mohlo být následkem toho, že v hokeji se vyskytuje stav remízy mnohem méně než ve fotbalovém odvětví.



Obrázek 15 - Ukázka grafu úspěšnosti predikce v třístavovém modelu (vítězství, prohra a remíza)



Obrázek 16 - Ukázka grafu úspěšnosti predikce v modelu, kdy je uvažován stav neprohry

8.3 Diskuze výsledků a použití PageRanku v tenise

Při zkoumání odvětví tenisu se zaměříme na hlavní mužské turnaje v singlu a v doublu (viz kapitola 3.3).

8.3.1 Výsledky úspěšnosti predikce v případě kladné predikce při úvaze třístavového modelu (vítězství, prohra a remíza)

V tomto odvětví byly očekávány nejlepší výsledky, jelikož tento sport je zaměřený na jednotlivce oproti oběma předcházejícím odvětvím, která jsou zaměřena na týmové výkony.

Výsledky Singlu mužských Grandslamových turnajů a Turnaje mistrů

Výsledky predikce pro tenisové odvětví, jsou nadprůměrné. V každém typu predikce výsledná hodnota průměrné úspěšnosti predikce pro celou soutěž přesáhla 55% hranici. Personalizace vrcholů je v tomto odvětví dána postavením vrcholů, představující hráče, v žebříčku ATP[5]. Čím výše postavený hráč v žebříčku je, tím vyšší hodnota personalizace je danému vrcholu přidělována.

[5] ATP – asociace tenisových profesionálů

Nejnižší výsledná hodnota byla zaznamenána v případě použití predikce u typu grafu, kde jsou váhy hran ohodnocené podle toho, jestli nastalo vítězství, prohra nebo remíza a to bez použití personalizace vrcholů hodnota byla 57,27%. I v tomto odvětví u tohoto typu grafu, hrálo použití personalizace vrcholů kladnou roli. Hodnota se při použití personalizace pro tento typ grafu zvýšila na 64,34%.

U druhého typu z grafů, kde jsou vahami hran získané sety danými vrcholy, byla výsledná hodnota průměrné úspěšnosti predikce ještě o trochu vyšší než v předešlém případě. Nicméně zde nastala situace, kdy použití personalizace hrálo zápornou roli v průměrné úspěšnosti predikce. Tedy u predikce bez využití personalizace vrcholů byla výsledná hodnota průměrné úspěšnosti predikce 67,15% a při využití personalizace vrcholů se daná hodnota snížila na 64,96%. Důsledkem by mohlo být to, že dané vrcholy u toho typu grafu nemohou získat více setů než 3, jelikož tenisové zápasy jsou obvykle hrány na 3 vítězné sety a proto není takový rozdíl mezi vahami hran.

Při použití typu grafu, kde jsou váhy hran určeny podle rozdílu výsledného skóre, dosahovala úspěšnost predikce také vyšších hodnot, než při zkoumání grafu s vahami hran podle vítězství, prohry nebo remízy, stejně jako u předchozího typu grafu, kde jsou vahami hran získané sety jednotlivými tenisty. Bet využitím personalizace vrcholů byla průměrná úspěšnost predikce 60,47% v případě využití personalizace vrcholů, se průměrná úspěšnost predikce zvýšila až na 69,88%, což byla nejvyšší hodnota úspěšnosti predikce pro danou soutěž.

Při predikci použitím tabulky byla úspěšnost predikce nejnižší, z čehož vyplívá, že tento způsob predikce není v tomto odvětví příliš dobrým řešením. Důvodem může být to, že zde není žádná tabulka pořadí hráčů, pouze žebříček ATP a navíc se tato soutěž hraje turnajově, kde nevždy musí být jedni a ti samý hráči. Tento fakt by mohl způsobit to, že daná hodnota průměrné predikce je u tohoto typu predikce nižší, než u předešlých typů. U tohoto typu predikce nabývá hodnoty 59,53%.

Výsledky Doublu mužských Grandslamových turnajů a Turnaje mistrů

V této soutěži byla nejhorší opět, stejně jako v předchozí soutěži, úspěšnost predikce v případě predikce podle aktuálního postavení v tabulce. Tento typ predikce v tomto odvětví nemá nejlepší úspěch, mohlo by to být způsobeno tím, že se nehraje na žádné body ani tabulky, jelikož by to ani nešlo, důvodem je to, že soutěž se hraje turnajovým způsobem a na každý turnaj se mohou kvalifikovat jiní hráči. Hodnota průměrné predikce u tohoto typu predikce byla rovna 45,37%.

Při použití typu grafu, kde jsou váhy hran určeny podle vítězství, prohry a remízy bez personalizace vrcholů se úspěšnost predikce, oproti předchozímu typu predikce, zvýšila na 47.84%.V totožném typu grafu vede využití personalizace vrcholů k lepším výsledkům predikce. Průměrná úspěšnost predikce je v tomto případě 59,81%.

U typu grafu, kde jsou vahami hran vyhrané sety je úspěšnost predikce s využitím personalizace vrcholů rovna 54,63%. Pokud není využita personalizace vrcholů, tak se úspěšnost predikce příliš neliší. Úspěšnost predikce je v tomto případe, bez využití personalizace vrcholů, rovna 52,48%.

Nejlepší úspěšnost predikce byla u typu grafu, kde jsou váhy hran určeny podle rozdílu výsledného skóre, ale pouze s využitím personalizace, která u toho typu grafu hraje velkou roli v úspěšnosti predikce. Hodnota průměrné úspěšnosti predikce byla v tomto případě rovna 60,35%, což jak už bylo zmíněno, byla nejlepší úspěšnost predikce pro tuto soutěž. Pokud není u tohoto typu grafu využito personalizace vrcholů, tak úspěšnost predikce klesá až na 49,13%.

Porovnání výsledků predikce obou zkoumaných soutěží tohoto sportovního odvětví

V tomto sportovním odvětví byl nejlepší výsledek predikce, v obou zkoumaných soutěžích, získán v případě využití predikce pomocí PageRanku společně s personalizací vrcholů pro graf, kde jsou váhy hran určeny podle rozdílu výsledného skóre. Z toho lze vypozorovat, že predikce je o něco úspěšnější pokud se zaměříme na konkrétní rozdíl získaných setů nebo jejich počet a ne pouze na vítězství či prohru, protože remíza v tenise nastat nemůže. Může nastat pouze tzv. skreč[6], poté hráč, který vyhrál po tom, co se soupeř vzdal, získává počet doposud získaných setů. To je zřejmě důvodem, proč je v tomto odvětví predikce využitím tohoto grafu úspěšnější než predikce při použití grafu, kde jsou váhy hran určeny podle vítězství, prohry a remízy.

8.3.2 Diskuze výsledků úspěšnosti predikce v případě kladné predikce při úvaze stavu neprohry

V případě tohoto sportovního odvětví žádná změna nenastane (viz Obrázek 17 a Obrázek 18), jelikož v tenise nemůže dojít k remíze.

Obrázek 17 - Ukázka grafu úspěšnosti predikce v třístavovém modelu (vítězství, prohra a remíza)

[6] Skreč – jeden z hráčů vzdá utkání

Obrázek 18 -Ukázka grafu úspěšnosti predikce v modelu, kdy je uvažován stav neprohry

8.4 Diskuze výsledků a použití PageRanku v házenkářských soutěžích

U házené se konkrétně zaměříme, na německou *Bundesligu* a na házenkářskou *Ligu mistrů* (viz část 4.4).

8.4.1 Výsledky úspěšnosti predikce v případě kladné predikce při úvaze třístavového modelu (vítězství, prohra a remíza)

U tohoto odvětví byly očekávány lepší výsledky predikce u typu grafů, kde jsou váhy hran určeny podle vstřelených gólů, jelikož v házené většinou padá 20 a více gólů za jeden zápas. Z toho důvodu by mohla predikce právě pro tento typ grafu přinést lepší výsledky než ostatní grafy a zároveň ostatní odvětví sportu.

Výsledky německé *Bundesligy*

V této soutěži byly získány nejhorší výsledky predikce v případě grafu, kde jsou vahami hran vstřelené góly, i když bylo očekáváno, že právě díky velkému počtu střílených gólů bude právě tento typ grafu, vykazovat nejlepší úspěšnost predikce. Pro tento typ grafu byla průměrná úspěšnost predikce rovna 57,84% bez využití personalizace vrcholů. Při zohlednění personalizace vrcholů vzrostla úspěšnost predikce na 67,65%, což je docela velký vliv na úspěšnost predikce s ohledem na jiná odvětví, kde rozdíl ve využití a nevyužití personalizace nebyl tak velký.

Pro typ grafu, kde jsou váhy hran určeny podle vítězství, prohry a remízy také nebyly výsledky úspěšnosti predikce nijak oslnivé. V případě personalizace vrcholů byla průměrná hodnota úspěšnosti predikce 63,07%. Nevyužití personalizace vrcholů mělo záporný dopad na změně hodnoty průměrné úspěšnosti predikce, která v tomto případě byla rovna 59,80%. Neleze tedy říci, že by využití tohoto grafu poskytlo lepší či horší úspěšnost predikce než jaká byla poskytnuta předchozím typem grafu, i když menší rozdíly vidět jsou.

Typ grafu, jehož váhy hran jsou určeny podle rozdílu výsledného skóre, poskytl při využití personalizace vrcholů nejlepší výsledek úspěšnosti predikce pro tuto soutěž. Tom je zřejmě způsobeno tím, že pokud nastane rozdíl ve výsledném skóre, tak je mnohem větší než např. ve

fotbalových soutěžím. Důvodem je počet střílených gólů, např. výsledek zápasu v házené může skončit např. výsledkem 32:25, z čeho vyplívá, že váha hrany bude rovna 7. Kdežto ve fotbalové zápase končí utkání většinou jednobrankovým nebo dvoubrankovým rozdílem, zřídkakdy nastane situace, že by rozdíl byl větší. Hodnota průměrné úspěšnosti predikce v případě tohoto grafu s využitím personalizace vrcholů je 69,88%. Pokud personalizace vrcholů není využito, klesá úspěšnost predikce na 60,78%.

Pokud jsou vítězové predikováni pouze podle aktuálního postavení v tabulce, tak je úspěšnost predikce tímto způsobem rovna 61,76%, z čeho lze vypozorovat, že predikování vítězů podle aktuálního postavení v tabulce není v této soutěži nejlepší řešením.

Výsledky házenkářské *Ligy mistrů*

Výsledky predikce v této soutěži byly horší než výsledky predikce pro soutěž předcházející. Pro tuto soutěž byly nejlepší výsledky predikce získány v případě typu grafu, kde jsou vahami hran vstřelené góly, a je využito personalizace vrcholů. Úspěšnost predikce byla rovna 61,98%. Zároveň u tohoto typu grafu nastala i nejhorší úspěšnost predikce pro tuto soutěž, a to při nevyužití personalizace vrcholů, úspěšnost predikce byla 52,34%.

U další dvou typů grafů byly výsledky celkem podobné, jako u většiny zkoumaných soutěží byla vždy nižší úspěšnost predikce při nevyužití personalizace vrcholů a vyšší, když je personalizace vrcholů využito.

U typu grafu, kde jsou váhy hran určeny podle vítězství, prohry a remízy je úspěšnost predikce bez využití personalizace rovna 54,43%, když je personalizace využito, tak se úspěšnost predikce zvyšuje na 60,16%.

Pro třetí typ grafu, kde jsou váhy hran určeny podle rozdílu výsledného skóre je rozdíl mezi úspěšností predikce pro tuto soutěž největší. S využitím personalizace vrcholů je úspěšnost predikce 59,90% a bez využití personalizace vrcholů se hodnota úspěšnosti predikce snižuje ne 55,21%.

V případě predikce využitím aktuálního postavení v tabulce je úspěšnost predikce podobná úspěšnostem předchozích typů grafů, kde je využito personalizace. Hodnota průměrné úspěšnosti predikce pro tento typ predikce je 60,16%.

Tyto výsledky predikce, které nejsou příliš dobré, může mít za následek počet kol této soutěže, kterých je pouze 16. Z toho důvodů není v grafu tolik hran, aby bylo možné dostatečně zvýhodnit určitě vrcholy.

Porovnání výsledků predikce obou zkoumaných soutěží tohoto sportovního odvětví

Výsledky predikce byly rozdílné v průměru o 5 procentuálních bodů. Rozdíl může spočívat v rozdílném počtu kol obou soutěží, když házenkářská *Liga mistrů* má pouze 16 kol a druhá zkoumaná soutěž, kterou je německá *Bundesliga*, má 34 kol. Rozdíl je tedy více než dvojnásobný, což by tedy mohlo být důsledkem rozdílných výsledků predikce.

8.4.2 Diskuze výsledků úspěšnosti predikce v případě kladné predikce při úvaze stavu neprohry

V případě tohoto sportovního odvětví je odlišnost daných modelů podobná hokejovým soutěžím, kde se v průměru liší o 11 procentuálních bodů (viz Obrázek 19 a Obrázek 20), což by mohlo být následkem toho, že v házené se, stejně jako v hokeji, tolik nevyskytuje stav remízy, ale určitý počet zápasů tímto výsledkem končí, proto zde nastává změna.

Obrázek 19 - Ukázka grafu úspěšnosti predikce v třístavovém modelu (vítězství, prohra a remíza)

Obrázek 20 - Ukázka grafu úspěšnosti predikce v modelu, kdy je uvažován stav neprohry

8.5 Diskuze výsledků a použití PageRanku v basketbalových soutěžích

U basketbalu se konkrétně zaměříme, na českou *Mattoni NBL* a na evropskou soutěž *Euroligu* (viz část 4.5).

8.5.1 Výsledky úspěšnosti predikce v případě kladné predikce při úvaze třístavového modelu (vítězství, prohra a remíza)

V tomto odvětví byly očekávány lepší výsledky predikce u typu grafů, kde jsou váhy hran určeny podle rozdílu výsledného skóre, jelikož v basketbalu jsou rozdíly výsledných skóre

většinou 10 a více bodů. Z toho důvodu by mohla predikce právě pro tento typ grafu přinést lepší výsledky než ostatní grafy a zároveň ostatní sportovní odvětví.

Výsledky *Euroligy*

V této soutěži se potvrdilo to, že zde budou vykazovány lepší výsledky predikce u typu grafu, kde jsou vahami hran získané body, podle počtu nastřílených košů, a u typu grafu, kde jsou váhy hran určeny podle rozdílu výsledného skóre.

Nejhorší výsledky predikce byly získány při zkoumání typu grafu, kde jsou váhy hran určeny podle vítězství, prohry a remízy. Bez personalizace vrcholů byla úspěšnost predikce 54,30% a s personalizací vrcholů se úspěšnost predikce zvýšila, tak jako ve všech zkoumaných soutěžích, na 56,59%.

Výsledky predikce získané u typu grafu, kde jsou vahami hran získané body, podle počtu vstřelených košů, bez využití personalizace vrcholů byla průměrná úspěšnost predikce rovna 55,91%, i zde se úspěšnost predikce zvýšila využitím personalizace, a to o 4,3 procentuální body na 60,21%. Je vidět, že při zaměření na počet získaných bodů, podle vstřelených košů a využitím personalizace vrcholů se úspěšnost predikce zvyšuje.

Nejlepší výsledky predikce ovšem nastaly u typu grafu, kde jsou váhy hran určeny podle rozdílu výsledného skóre, tak jak bylo u této soutěže předpokládáno. Nicméně rozdíl úspěšnosti predikce není tak velký, jak se čekalo. Úspěšnost predikce u tohoto typu grafu s využitím personalizace vrcholů byla u tohoto typu grafu 62,37%, v případě nevyužití personalizace je úspěšnost predikce o něco nižší a to 60,22%.

Při predikci podle aktuálního postavení v tabulce byla hodnota průměrné úspěšnosti predikce rovna 60,75%. Z toho lze vyvodit, že pro tuto soutěž byla predikce podle aktuálního postavení v tabulce druhou nejlépe úspěšnou metodou predikce.

Výsledky *Mattoni NBL*

V této soutěži je nejhorší úspěšnost opět v případě typu grafu, kde jsou váhy hran určeny podle vítězství, prohry a remízy, stejně jako v předchozí soutěži. Úspěšnost predikce bez využití je 52,87%, nicméně při využití personalizace vrcholů, se úspěšnost predikce v této soutěži zvýšila na 61,96%. Jedná se o celkem znatelné rozdíl při predikci s využitím personalizace vrcholů, oproti předchozí soutěži. Což může být spojeno s rozdílnými strukturami a průběhy daných soutěží. V *Eurolize* se navíc hraje méně zápasů než v *Mattoni NBL*.

V případě zkoumání typu grafu, kde jsou váhy hran rovné počtu získaných bodů, podle vstřelených košů, bez využití personalizace vrcholů, je průměrná úspěšnost predikce 57,95%. Využití personalizace vrcholů hraje v tomto případě zápornou roli při úspěšnosti predikce, i když rozdíl je zanedbatelný. Hodnota průměrné úspěšnosti predikce při využití personalizace vrcholů sníží o 0,92 procentuálních bodů na 57,03

Nejlepší výsledky predikce byly získány opět, stejně jako v předchozí soutěži, z typu grafu, kde jsou váhy hran určeny podle rozdílu výsledného skóre. Je tedy opět potvrzen předpoklad, že právě u tohoto typu grafu s využitím personalizace vrcholů je úspěšnost predikce jednou z

nejlepší, konkrétně 66,46%. V této soutěži, u tohoto typu grafu, ale hraje, oproti ostatním soutěžím, personalizace vrcholů zápornou roli. Pokud tedy není personalizace vrcholů u tohoto typu grafu využita, tak se průměrná úspěšnost predikce zvýší na 67,63%, což je nejvyšší úspěšnost predikce pro tuto soutěž. Důvodem by mohlo být např. zranění klíčových hráčů, pokles formy atd.

Úspěšnost predikce podle aktuálního postavení v tabulce je v této soutěži, s ohledem na ostatní typy predikce, průměrná. Hodnota průměrné úspěšnosti predikce je 62,81%. Z čeho lze usoudit, že v této soutěži je využití tohoto typu predikce lepší, než predikce podle PageRanku pro jakýkoliv graf bez personalizace vrcholů.

Porovnání výsledků predikce obou zkoumaných soutěží tohoto sportovního odvětví

U obou zkoumaných soutěží byly získané výsledky predikce podobné. Nejlepší úspěšnost predikce nastala v případě typu grafu, kde jsou váhy hran určeny podle rozdílu výsledného skóre. Což bylo pro tuto soutěž předpokládáno. V ostatních typech predikce se úspěšnost predikce v obou soutěžích příliš nelišila

8.5.1 Diskuze výsledků úspěšnosti predikce v případě kladné predikce při úvaze stavu neprohry

V případě *Euroligy* jsou výsledky průměrně predikce rozdílné, s ohledem dvou daných modelů, o 6 procentuálních bodů. V *Mattoni NBL* je úspěšnost průměrně predikce rozdílná o pouhé 2 procentuální body. Rozdíl obou soutěží by měl být způsoben tím, že v *Eurolize* nastalo ve zkoumaném ročníku více remíz než v *Mattoni NBL*.

8.6 Porovnání výsledků a použití PageRanku pro všechna zkoumaná odvětví

8.6.1 Porovnání výsledků predikce v třístavovém modelu (vítězství, prohra a remíza) a v modelu, kde je uvažován stav neprohry (viz část 7.2)

Při porovnání výsledků úspěšné predikce při využití třístavového modelu (výhra, prohra a remíza) a modelu, kde je uvažováno stavu neprohry je logickým zjištěním, že výsledné hodnoty úspěšnosti predikce se při úvaze stavu neprohry zvýší. Důvod je logický, protože predikce vítěze je úspěšná i v případě, kdy nastane remíza.

8.6.2 Diskuze výsledků a použití PageRanku pro všechna zkoumaná odvětví

Výsledky nejsou nijak oslnivé. Bylo předpokládáno, že nejlepší výsledky by mělo vykazovat odvětví tenisu, jelikož tenis je individuálním sportem. Nemusí se tedy zohledňovat např. zranění klíčového hráče a tak podobně.

Tento předpoklad se potvrdil, v tenise byly opravdu získány nejlepší výsledky ze všech zkoumaných sportů. Nicméně pouze v oblasti tzv. singlu, tedy soutěže jednotlivců Také bylo zjištěno, že přeci jen o trochu lepším řešením bylo využití typu grafu, kde jsou hodnoty vah hran určeny podle vítězství, prohry a remízy s využitím personalizace vrcholů, která vždy (až

na čtyři případy, viz části 8.3.1.1, 8.4.1.1 a 8.5.1.2) zvyšovala úspěšnost predikce. V tenise byla úspěšnost v prvním kole nulová (pokud nebylo využito personalizace vrcholů), jelikož v tenise nemůže nastat remíza, ale aplikace predikuje v prvním kole pouze remízu, protože všichni mají buď stejnou hodnotu PageRanku (pokud není využito personalizace vrcholů) nebo stejný počet bodů v tabulce. Ale s přibývajícími koly se úspěšnost predikce zlepšuje.

Oproti tomu nejhorší výsledky úspěšnosti predikce nastali v odvětví fotbalu, konkrétně v *I.A třídě Karlovarského kraje*, kde chyby v predikci vítěze mohou nastat, zejména protože tato soutěž není vyrovnaná a je možné, že dané zápasy jsou ovlivňovány od zástupců týmu, kteří chtějí pomoci „své věci“.

Uspokojivé výsledky úspěšnosti predikce nebyli ani v odvětví hokeje. V tomto odvětví jsme předpokládaly dobré výsledky, vzhledem k velkému množství interakce mezi vrcholy grafu, která je reprezentována tvořenými hranami podle skutečných výsledků sledovaných zápasů.

Lepší výsledky predikce vykazovali soutěže, kde padá hodně gólů a výsledné skoré je vysoké (házená, viz část 8.4 a basketbal, viz část 8.5). Pokud byly v těchto soutěžích zkoumány typy grafů, kde jsou vahami hran vstřelené góly (házená) nebo získané body (basketbal), nebo jsou hrany určeny podle rozdílu výsledného skóre, tak se úspěšnost predikce zvyšovala. Dobré výsledky vykazovaly tyto soutěže, protože vítězný vrchol je více zvýhodněn v případě, když do něj vstupuje hrana s vahou např. 84 (počet bodů získaných v basketbale) než v případě hrany s vahou např. 3 (počet vítězných setů v tenise).

I když nebyly získány příliš dobré výsledky, tak se v každé soutěži s přibývajícími koly úspěšnost predikce zvyšovala, ale několikrát nastala i situace, kdy, např. v předposledním kole, nebyla ani jedna z predikcí úspěšná.

Ze získaných výsledků můžeme konstatovat, že pokud při uvažování stavu neprohry, se výsledky úspěšnosti predikce lepší, v některých soutěžích až o 25 procentuálních bodů, pouze v tenise zůstávají stejné, což způsobeno tím, že v tenise nemůže nastat remíza. Po analýze výsledků je využití PageRanku pro predikci vítězů v oblasti sportovních utkání minimální.

9 Závěr

Cílem práce bylo seznámit čtenáře s používanými metodami pro predikci budoucí hrany v sociální nebo jiné síti. Dále s používaným algoritmem PageRank. Následovalo určení sportovních odvětví a jejich konkrétních soutěží, na kterých bylo využití algoritmu PageRank pro predikci vítězů testováno.

V teoretické části byly popsány metody pro predikci budoucí hrany v sociální a jiných sítích (dále jen graf). V další části byl popsán algoritmus PageRank, byly zmíněny jeho problémy a jejich možné řešení, což vedlo ke zdárné implementaci algoritmu. Dále byly navrženy 2 typy grafů a úprava personalizačního vektoru, na kterých byla testována úspěšnost predikce budoucí hrany v této oblasti.

V praktické části byly vytvořeny vstupní soubory dané pevné struktury, tak aby byly vhodné pro vytvořenou aplikaci. Následně byla vytvořena aplikace, která slouží k načtení dat ze vstupního souboru, postupnému vytváření grafu a následné predikci vítězů sledovaných sportovních utkání, podle daných typů predikce. Výsledky jsou zapisovány do výstupního souboru.

Pro lepší prezentaci a analýzu výsledků byla vytvořena další aplikace, která slouží k vykreslování jednoduchého grafu, kde je procentuálně znázorněna úspěšnost predikce pro jednotlivá kola zkoumané soutěže. Pro lepší porovnání s ostatními typy predikce a hlavně ostatními soutěžemi obsahuje vykreslený graf hodnotu průměrné úspěšnosti predikce pro celou soutěž.

Ve výsledcích průměrné úspěšnosti predikce vítězů pro celou soutěž nebyly s ohledem na soutěže nijak velké rozdíly, spíše se jednalo o malé odchylky. Pouze v tenise, byli výsledky výraznějším způsobem lepší než v ostatních sportovních odvětvích a to zejména, protože se jedná o sport jednotlivců oproti ostatním odvětvím, kde jde o kolektivní sporty.

Ze získaných výsledků je možné soudit, že daný způsob predikce budoucí hrany v této oblasti nenajde široké využití, jelikož výsledky se pohybovali okolo průměrných 50% hodnot. Což nejsou nijak skvělé hodnoty úspěšnosti predikce budoucí hrany v grafu.

Literatura

[1] R. Albert, A.-L. Barabási. *Statistical mechanics of complex network*. Rev. Modern Phys. 74 (2002) 47.

[2] S. N. Dorogovtsev, J.F.F. Mendes. *Evolution of network*. Adv. Phys. 51 (2002) 1079.

[3] L. Getoor, C. P. Dieh., Link mining: a survey. ACM SIGKDD Explor. Newsl. 7 (2005) 3.

[4] S. Zhou, R. J. Mondragón. *Accurately modeling the internet topology*. Phys. Rev. E 70 (2004) 066108.

[5] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, E. Shir, *A model of Internet topology using k-shell decompositio.* Proc. Natl. Acad. Sci. USA 104 (2007) 11150.

[6] M. Sales-Pardo, R. Guimerà, L.A.N. Amaral. *Extracting the hierarchical organization of complex systems*. Proc. Natl. Acad. Sci. USA 104 (2007) 15224.

[7] M. Girvan, M. E. J. Newman. *Community structure in social and biological networks*. Proc. Natl. Acad. Sci. USA 99 (2002) 7821.

[8] Z. Huang, X. Li, H. Chen. *Link prediction approach to collaborative filtering*. in: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, ACM Press, New York, 2005.

[9] D. Lin. *An information-theoretic definition of similarity*. in: Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufman Publishers, San Francisco, 1998.

[10] D. R. White, K. P. Reitz. *Graph and semigroup homomorphisms on networks of relations*. Soc. Networks 5 (1983) 193.

[11] P. Holme, M. Huss. *Role-similarity based functional prediction in networked systems: application to the yeast proteome.* J. R. Soc. Interface 2 (2005) 327.

[12] S. Redner. *Teasing out the missing links*. Nature 453 (2008) 47.

[13] H. C. White, S. A. Boorman, R. L. Breiger. *Social structure from multiple networks I: blockmodels of roles and positions*. Am. J. Sociol. 81 (1976) 730.

[14] P. W. Holland, K. B. Laskey, S. Leinhardt. *Stochastic blockmodels: first step*. Soc. Networks 5 (1983) 109.

[15] P. Dorelan, V. Batagelj, A. Ferligoj. *Generalized Blockmodeling*. Cambridge University Press, Cambridge, UK, 2005.

[16] E. M. Airoldi, D. M. Blei, S. E. Fienberg, X. P. Xing. *Mixed-membership stochastic blockmodels*. J. Mach. Learn. Res. 9 (2008) 1981.

[17] W. Zachary. *An information flow model for conflict and fission in small Gross*. J. Anthropol. Res. 33 (1977) 452.

[18] D. Lusseau, et al.. *The bottlenose dolphin community of Doubtful sound features a large proportion of long-lasting associations*. Behav. Ecol. Sociobiol. 54 (2003) 396.

[19] R. Guimerà, S. Mossa, A. Turtschi, L.A.N. Amaral. *The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles*. Proc. Natl. Acad. Sci. USA 102 (2005) 7794.

[20] J. G. White, E. Southgate, J. N. Thomson, S. Brenner. *The structure of the nervous system of the nematode C. elegant*. Philos. Trans. R. Soc. Lond. Ser. B 314 (1986) 1.

[21] J.L. Reed, T. D. Vo, C.H. Schilling, B.Ø Palsson. *An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR)*. Genome Biol. 4 (2003) R54.

[22] R. Guimerà, M. Sales-Pardo. *Missing and spurious interactions and the reconstruction of complex networks*. Proc. Natl. Acad. Sci. USA 106 (2009) 22073.

[23] N. Friedman, L. Getoor, D. Koller, A. Pfeffer. *Learning probabilistic relational models*. in: Proceedings of the 16th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 1999, p. 1300.

[24] D. Heckerman, C. Meek, D. Koller. *Probabilistic entity-relationship models, PRMS, and plate models*. in: Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004, p. 55.

[25] K. Yu, W. Chu, S. Yu, V. Tresp, Z. Xu. *Stochastic relational models for discriminative link prediction*. in: Proceedings of Neural Information Precessing Systems, MIT Press, Cambridge, MA, 2007, pp. 1553–1560.

[26] A. Clauset, C. Moore, M. E. J. Newman. *Hierarchical structure and the prediction of missing links in networks*. Nature 453 (2008) 98.

[27] J. O'Madadhain, J. Hutchins, P. Smyth. *Prediction and ranking algorithms for event-based network data*. in: Proceedings of SIGKDD 2005, ACM Press, New York, 2005, p. 23.

[28] D. Liben-Nowell, J. Kleinberg. *The link-prediction problem for social networks*. J. Am. Soc. Inf. Sci. Technol. 58 (2007) 1019.

[29] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. (S. University, Ed.)Computer Networks and ISDN Systems, 30(1-7), 107–117. doi:10.1016/S0169-7552(98)00110-X

[30] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999, November 11). The PageRank Citation Ranking: Bringing Order to the Web. Stanford InfoLab. Převzato z *http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf*

[31] Nykl, M.: Určování významnosti vrcholů grafu: PageRank a jeho modifikace. Technical report No. DCSE/TR-2013-09, University of West Bohemia, 2013.

Obsah CD

Struktura a obsah přiloženého DVD:

- **Vstupní soubory**
  - Basket
    - BasketEuroliga.csv – vstupní soubor pro baskebalovou Euroligu
    - MattoniNBL.csv – vstupní soubor pro českou Mattoni NBL
  - Házená
    - HazenaLigaMistru.csv – vstupní soubor pro házenkářskou Ligu mistrů
    - HazenaBundesliga.csv – vstupní soubor pro německou Bundesligu
  - Hokej
    - Extraliga.csv – vstupní soubor pro českou hokejovou extraligu
    - NHL.csv – vstupní soubor pro NHL
  - Fotbal
    - PremierLeague – vstupní soubor pro anglickou Premier League
    - I.ATrida.csv – vstupní soubor pro I.A třídu karlovarského kraje
    - LigaMistru.csv – vstupní soubor pro fotbalovou Ligu mistrů
  - Tenis
    - Tenis.csv – vstupní soubor pro soutěž singlistů (jednotlivců)
    - TenisCtyrhra.csv – vstupní soubor pro soutěž dvojic (double)
- **Výtup**
  - Basket
    - Obsahuje výstupní soubory pro všechny typy grafů i predikce a vykreslené grafy k těmto výstupním souborů pro basketbalové soutěže
  - Házená
    - Obsahuje výstupní soubory pro všechny typy grafů i predikce a vykreslené grafy k těmto výstupním souborů pro házenkářské soutěže
  - Hokej
    - Obsahuje výstupní soubory pro všechny typy grafů i predikce a vykreslené grafy k těmto výstupním souborů pro hokejové soutěže
  - Fotbal
    - Obsahuje výstupní soubory pro všechny typy grafů i predikce a vykreslené grafy k těmto výstupním souborů pro fotbalové soutěže
  - Tenis
    - Obsahuje výstupní soubory pro všechny typy grafů i predikce a vykreslené grafy k těmto výstupním souborů pro tenisové soutěže
- **Predikce.jar –** Aplikace pro tvorbu navržených typů grafů a jejich následného vyhodnocování PageRankem
- **Graf.jar** – Aplikace pro vykreslování grafů úspěšnosti predikce
- **BP_sudap_A11B0612P** – elektronikcá verze BP
- **BP_sudap.rar –** zdrojové dokumenty elektronické verze BP

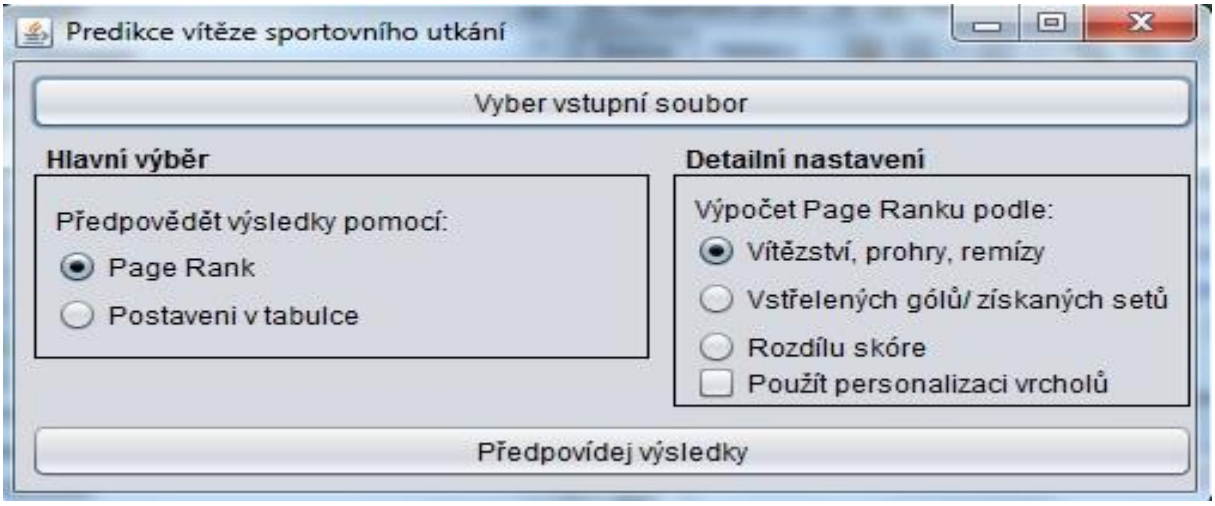
Přílohy

A Uživatelská příručka

A.1 Aplikace pro vytvoření (navržených) grafů a jejích vyhodnocení PageRankem

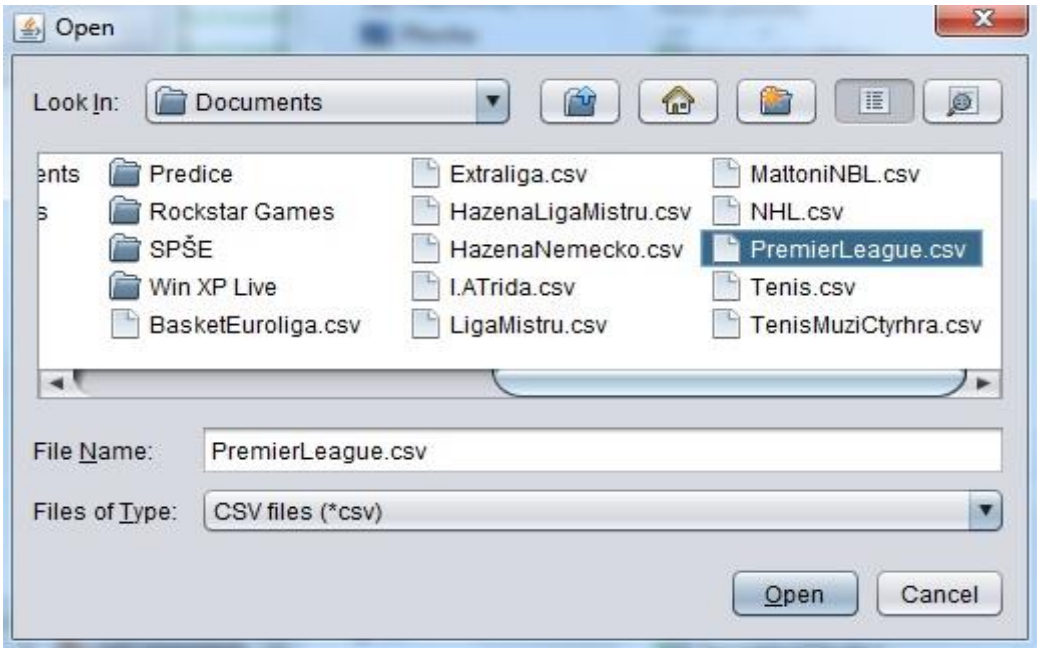
Aplikaci je možné spustit dvěma způsoby. První ze způsobů je poklepáním na *Predikce.jar*, druhým způsobem je spuštění aplikace pomocí příkazové řádky a to tím způsobem, že se přepneme do adresáře, kde se daný soubor *Predikce.jar* nachází a tomto adresáři provedeme příkaz: *java –jar Predikce.jar*.

Po spuštění jedním ze způsobů se zobrazí hlavní dialogové okno (viz Obrázek 21). Kde se při kliknutí na tlačítko *Vyber vstupní soubor* zobrazí formulář pro výběr vstupního souboru (viz Obrázek 22), který musí mít pevnou strukturu a být typu CSV (oddělený středníkem) (viz část 6.2). Po výběru vhodného souboru, si v hlavním dialogovém okně můžeme vybrat, podle čeho chce vítěze predikovat. Při predikci využitím PageRanku, máme k dispozici detailní nastavení, kde si vybereme typ grafu (viz kapitola 5), který chceme využít a jestli chceme využít personalizace vrcholů nebo ne. V případě, že vstupní soubor neobsahuje informace o personalizaci vrcholů, je vypsána hláška o dané skutečnosti. Pokud chceme predikovat vítěze podle aktuálního postavení v tabulce, není detailní nastavení k dispozici. Po kliknutí na tlačítko *Předpovídej výsledky* se zobrazí formulář uložení výstupního souboru (viz Obrázek 23), který bude obsahovat dané výsledky (viz část 6.4). Po zadání názvu výstupního souboru aplikace začne predikovat vítěze daných sportovních utkání a výsledky zapije do nově vytvořeného vstupního souboru s názvem, který uživatel zvolil.

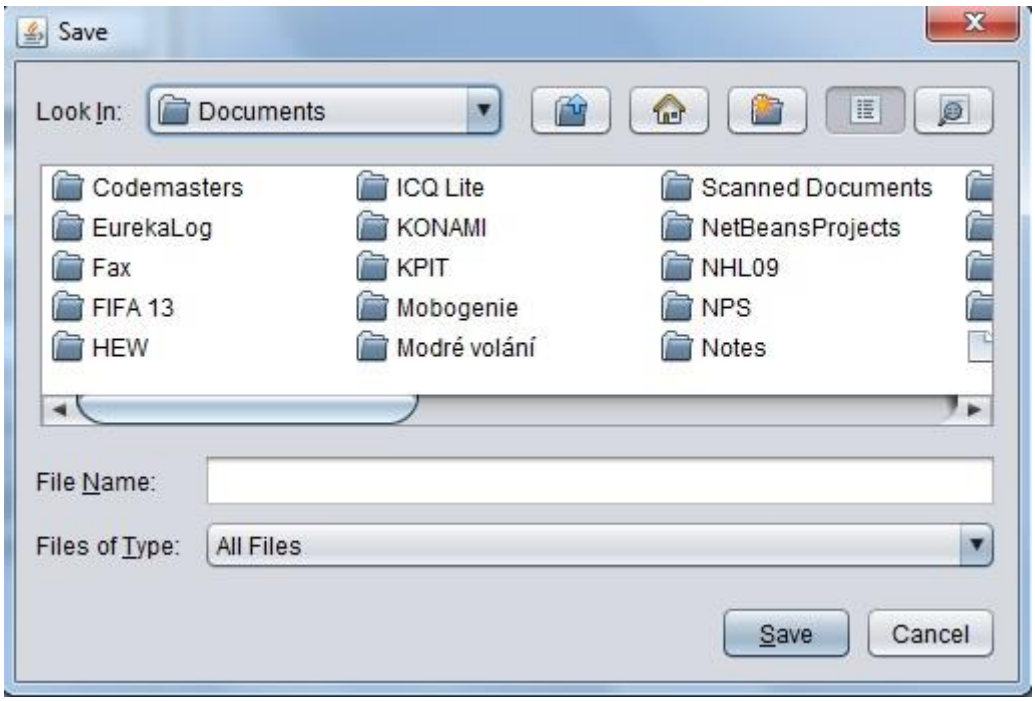


Obrázek 21 - Hlavní okno aplikace

Ukončení aplikace se provádí pomocí křížku v pravém horním rohu hlavní okna.



Obrázek 22 - Formulář pro načtení vstupního souboru



Obrázek 23 - Formulář pro uložení výstupního souboru

A.2 Aplikace pro vykreslení grafu úspěšnosti predikce

Aplikaci je možné spustit stejnými způsoby jako aplikaci pro vytvoření (navržených) grafů a jejich následného vyhodnocení PageRankem. Pouze je v obou případech nutné zaměnit soubor *Predikce.jar* za soubor *Graf.jar*.

Po spuštění aplikace jedním z uvedených způsobů, se zobrazí formulář pro vstupní soubor (viz Obrázek 13), který musí mít pevnou strukturu a být typu CSV (oddělený středníkem) (viz část 6.2). Po tom, co si uživatel vybere vstupní soubor, se zobrazí formulář (viz Obrázek 15),

kde je uvedena hodnota průměrné úspěšnosti predikce pro celou soutěž. Dále je v něm vykreslený graf, pomocí něhož je znázorněna úspěšnost predikce pro každé kolo soutěže. Formulář obsahuje zaškrtávací políčko, které slouží pro překreslení graf. Pokud bychom nechtěli uvažovat třístavový model (vítězství, prohra a remíza), ale model kde je uvažovaný stav neprohry (viz část 7.2), tak musíme dané zaškrtávací políčko označit (viz Obrázky 24 a Obrázek 25). Pokud bychom chtěli zpět původní graf, tak ho zase „odoznačit“. V poslední řadě obsahuje formulář tlačítko *Uložit graf jako JPEG*, po kliknutí na toto tlačítko se zobrazí formulář pro uložení grafu ve formátu JPEG (viz Obrázek 23). Aplikace se ukončuje prostřednictvím křížku v pravém horním rohu.

Obrázek 24 - Formulář zobrazující graf bez zaškrtnutého políčka pro překreslení

Obrázek 25 - Formulář zobrazující graf se zaškrtnutým políčkem pro překreslení, jak je vidět oba grafy jsou odlišné, jelikož spodní graf je překreslený